

For Reference

NOT TO BE TAKEN FROM THIS ROOM

Ex libris
UNIVERSITATIS
ALBERTAEASIS



THE EMPIRICAL BAYES PROBLEM

by

GERALD C. KOZUB

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF MASTER OF SCIENCE

DEPARTMENT OF MATHEMATICS

UNIVERSITY OF ALBERTA

EDMONTON, ALBERTA

MARCH, 1967

UNIVERSITY OF ALBERTA
FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read
and recommend to the Faculty of Graduate Studies for
acceptance, a thesis entitled "THE EMPIRICAL BAYES PROBLEM",
submitted by GERALD C. KOZUB in partial fulfillment of the
requirements for the degree of Master of Science.

ABSTRACT

Consider a statistical decision problem which involves an observable random variable X whose distribution, which depends on an unknown parameter λ , is known. Suppose that this problem occurs repeatedly. The sequence $\lambda_1, \lambda_2, \dots$ is regarded as the realization of the independent random variables $\Lambda_1, \Lambda_2, \dots$ with a common unknown a priori probability distribution. The "empirical Bayes problem" refers to the approximation of the Bayes decision function which would be optimal if the a priori distribution were known in advance.

Following an introduction and history of the empirical Bayes problem in Chapter I, five main areas are discussed. In Chapter II asymptotically optimal empirical Bayes estimators and tests of hypotheses concerning the true value λ of the random variable Λ are given. Two methods of estimating the a priori distribution are given in Chapter III.

In Chapter IV the empirical Bayes approach is used in selecting the "best" of k populations or a subset containing the "best". Empirical Bayes procedures are obtained in the case where the a priori distribution on the parameter space is assumed to belong to a particular parametric family and the case where the a priori distribution is assumed to possess a finite absolute mean.

Chapter V deals with the non-parametric case in which the class of conditional probability distributions of X given λ is not restricted to a particular parametric family. Analogous results to Chapter II are obtained, and an application is made to the problem of selecting the "best" of k populations.

The last chapter is concerned with the closely related compound decision problem. The unknown parameter values $\lambda_1, \lambda_2, \dots$ are regarded as a sequence of arbitrary unknown constants, and information is obtained from the observed sequence x_1, x_2, \dots regarding the frequency distribution of the parameters. The aim is to approximate the decision function which would be optimal if the frequency distribution of the parameters were known in advance.

ACKNOWLEDGEMENTS

The author would like to express his gratitude to Dr. K. L. Mehra for suggesting this topic and for his guidance. He would also like to thank Mr. George Smith for the many useful discussions, Professor E. S. Keeping and Dr. K. V. Wilson for reading the manuscript, and Miss June Talpash for doing the necessary typing.

TABLE OF CONTENTS

	Page
ABSTRACT	(i)
ACKNOWLEDGEMENTS	(iii)
CHAPTER I: HISTORY OF THE PROBLEM	1
CHAPTER II: PARAMETRIC EMPIRICAL BAYES PROBLEM FOR TESTING HYPOTHESES AND POINT ESTIMATION	
2.1 Introduction	8
2.2 General Results on Asymptotic Optimality	12
2.3 Application for Two Decision Problems	17
A. Discrete Case	17
B. Continuous Case	21
2.4 General Methods for Obtaining Consistent Estimators in the Two Decision Problem	24
A. General Method I	26
B. General Method II	32
2.5 Point Estimation	39
CHAPTER III: ESTIMATING THE PRIOR DISTRIBUTION	
3.1 Robbin's General Method	49
3.2 Squared Error Consistent Estimates of the Prior Distribution Function .	58

	Page
CHAPTER IV: SELECTING THE BEST OF k POPULATIONS PARAMETRIC CASE	
4.1 Introduction	64
4.2 Selecting a Subset Containing the Best Population	68
4.3 Bayes Procedures When G Belongs to a Particular Parametric Family . .	72
4.4 Empirical Bayes Procedures When G Belongs to a Particular Parametric Family	80
4.5 Examples	83
4.6 Empirical Bayes Procedures. The General Case	94
CHAPTER V : NON-PARAMETRIC EMPIRICAL BAYES ESTIMATION AND HYPOTHESIS TESTING	
5.1 Estimation. Discrete Case	101
5.2 Supplementary Sample Method . . .	105
5.3 Estimation. Continuous Case . . .	111
5.4 Application to Hypothesis Testing .	115
5.5 Non-parametric Empirical Bayes Approach For Selecting the Best of k Populations	121
CHAPTER VI: COMPOUND DECISION PROBLEM	131
BIBLIOGRAPHY	147

CHAPTER I

HISTORY OF THE PROBLEM

Consider a parameter space Ω such that to each parameter point $\lambda \in \Omega$ there corresponds a conditional probability distribution P_λ over some sample space E with corresponding conditional distribution function F_λ . An observable random variable X with values in E follows one of the distributions P_λ where λ is unknown. An appropriate experiment yields the observation $X = x$ from which we wish to infer the corresponding value η of the parameter λ before taking some contemplated action. If \mathcal{D} is the set of all possible actions, how can we best define a function $\delta(x)$ on E with values in \mathcal{D} which optimizes, in some sense, the selection of the action to be taken in accordance with the results of observations?

The following examples are applications of the above problem.

Example 1. Acceptance sampling. Consider lots containing N items. In order to decide whether a lot should be accepted or not, it is customary to select n items randomly from it for inspection. X is the number of defectives in the sample, and λ is the number of defectives in the lot. The parameter space Ω consists of the $N + 1$ integers $0, 1, 2, \dots, N$. The set \mathcal{D} of actions contemplated may include only two elements, "accept lot" and "reject lot".

Example 2. Diagnostic tests. It is common practice to assume that the seriousness of a disease can be expressed in terms of a parameter λ . The parameter λ cannot be measured directly for an individual, but a series of diagnostic tests are given which provide data for the estimation of λ . Thus the value of a random variable X , whose distribution depends on λ , is observed. Here again, λ has a certain range Ω of possible values, and, at least in some cases, the distribution of X given λ may be assumed to be known.

The "classical" solution to the above and similar problems is given by Bayes' formula, which requires that λ is the realization of a random variable Λ with a priori probability distribution $G(\lambda) = P[\Lambda \leq \lambda]$. Considering the overall sample space $\Omega \times E$, the random variables (Λ, X) are assumed to have a joint distribution. For given G and any observed $X = x$, Bayes' formula gives the conditional distribution of Λ given $X = x$, say

$$(1.1) \quad d B(\lambda|x) = \frac{f_{\lambda}(x) d G(\lambda)}{\int_{\Omega} f_{\lambda}(x) d G(\lambda)}$$

where $f_{\lambda}(x)$ is the conditional probability density of X given $\Lambda = \lambda$. λ may be estimated by computing the mean of the distribution (1.1).

Except in rare cases, Bayes' formula cannot be applied unless λ is the realization of a random variable Λ with an a priori distribution. The data in Example 2 do not contain anything regarding this distribution,

and in Example 1, λ may not be regarded as the realization of a random variable.

The problem thus arose of devising a mathematical theory of using the observable X in order to provide a justifiable selection of decisions or actions regarding the true value λ of Λ , even in those cases where the datum of a problem does not include the a priori distribution $G(\lambda)$.

Early results by Bernstein and Von Mises (around 1920) stated generally that if we consider λ to be the realization of a random variable Λ with a continuous probability density, then as the number of conditionally independent observations is increased, the conditional distribution of Λ given $X = x$, given by (1.1), tends to a calculable limit, independent of the a priori distribution $G(\lambda)$. The requirement of a large number of observations was not practical in most cases, and in others there was a reluctance to consider λ as a realization of a random variable.

Other attempts to solve the problem, such as the "principle of insufficient reason", which states generally that if one is ignorant of the a priori distribution, one "has the right" to assume that it is uniform over Ω , were of little help. In the 1930's, the functions $\delta(x)$, now regarded as random variables, were the principal subject of study and were called statistical decision functions. Their properties were studied, and references were made to "losses" incurred due to "errors", the primary concern being the case where λ was not the realization of

a random variable.

After World War II there were protests that the use of previous experience, which would indicate that some values of the parameter λ are more probable than others, was being ignored by tests of hypotheses and confidence intervals. In 1955, Robbins [12] utilized previous experience in the following manner. It was assumed that λ is the realization of a random variable Λ with unknown distribution G and that the problem of estimating the true value λ of Λ from an observed value of X , a discrete random variable, occurred repeatedly. Thus there was a sequence (Λ_i, X_i) , $i = 1, 2, \dots, n$, where each Λ_i was assumed to have the same unknown marginal distribution G , and given $\Lambda_i = \lambda_i$, the corresponding random variable X_i was assumed to have the known conditional probability density $f_{\lambda_i}(x)$.

Assuming that the values $\lambda_1, \lambda_2, \dots, \lambda_n$ are all unobservable realizations of the random variables $\Lambda_1, \Lambda_2, \dots, \Lambda_n$ respectively, and that the observations are limited to the values of X_1, X_2, \dots, X_n , Robbins used all these observations and the postulated existence of G in order to estimate λ_n . This was called the empirical Bayes problem and corresponds exactly with the situation in Example 2.

The empirical Bayes procedure shows essentially that if the amount of previous experience is large, then the Bayes estimate of λ that could be computed with full knowledge of the a priori distribution of this parameter cannot be much better than Robbins' empirical Bayes estimate of λ . Robbins indicated the need for selecting the estimate

that was, in some sense, best.

Shortly after the initial breakthrough by Robbins, Johns [6], in 1957, studied the non-parametric empirical Bayes problem where the class of conditional probability distributions of the random variable X is not restricted to a particular parametric family. After a lull of several years, Robbins [13], [14] and Samuel [20] renewed work on the empirical Bayes problem in the early 1960's. Both discussed the empirical Bayes approach to statistical decision problems, and Robbins introduced methods of estimating the a priori distribution.

In recent years there has been a surge of activity in this area. Rutherford and Krutchkoff [16], [17] and [18] have extended results on the parametric empirical Bayes approach to statistical hypotheses testing, point estimation, and the estimation of the prior distribution. Krutchkoff [8] has discussed the non-parametric empirical Bayes approach to decision theory. Deely [2], [3] has applied the method to the problem of selecting the best of k populations in both the parametric and non-parametric cases.

Often, as in Example 1, there may be a reluctance to consider that the unobservable parameter λ is the realization of a random variable Λ with a fixed but unknown a priori distribution. Suppose we have N disconnected problems which are of the testing hypotheses or estimation type. In acceptance sampling, for example, there may be $N = 1000$ lots of shoes, each having λ defectives. The problem is to decide whether the i 'th lot should or should not be accepted as conforming with the

agreed specifications for $i = 1, 2, \dots, N$.

Initially it was thought that the N identical decision problems should be treated separately in the best possible way, perhaps by using most powerful tests. In 1950, Robbins [11] stated that if a large number N of identical but unrelated decision problems are treated simultaneously (the compound statistical decision problem), then in certain circumstances the overall expected frequency of errors of both kinds will be below the level attainable if N independent applications of the most powerful test were made. The requirement of simultaneity of N decisions was later dropped and an appropriate sequential procedure was substituted for the original solution.

In 1955 Hannan and Robbins [4] studied the nonsequential case where the component problems involve decisions between any two completely specified distributions. More recently, Samuel [19] and [21], Hannan and Van Ryzin [5], and Johns [7] have extended the early results of Hannan and Robbins for both the sequential and nonsequential cases. Van Ryzin [24] also considers the case where the component problem involves a finite parameter space and a finite action space. At the present time, work is also progressing on the standard (infinite state) compound estimation problem with notable results having already been achieved by Samuel [22] and Swain [23].

The compound decision problem appears to be related to the empirical Bayes problem, which uses the accumulated experience, in the following way. Suppose that before the N decisions have to be made,

the statistician knows the values of the parameters λ_i , $i = 1, 2, \dots, N$, but not their order; i.e. the function $G_N(\lambda) = \frac{1}{N}$ (number of indices i , $i = 1, 2, \dots, N$ for which $\lambda_i \leq \lambda$) is known. Then at each decision the statistician can use the rule which is Bayes with respect to G_N .

CHAPTER II

PARAMETRIC EMPIRICAL BAYES PROBLEM FOR TESTING

HYPOTHESES AND POINT ESTIMATION

2.1. Introduction.

Consider a statistical decision problem which involves an unknown real parameter λ belonging to some set Ω and an observable random variable X distributed according to the distribution function $F_\lambda(x) = F(x|\lambda) = P(X \leq x|\lambda)$ which is known for each $\lambda \in \Omega$. If we assume that λ is the realization of an unobservable random variable Λ with a known distribution function $G(\lambda)$ (the a priori distribution), then on the basis of observation $x \in \mathcal{X}$ of X we wish to estimate or make a test concerning the true value λ of Λ .

To do this, let $\mathcal{D} = \{d\}$ be the set of possible decisions. For example, \mathcal{D} may consist of two decisions in the case of testing a hypothesis, or decision $d \in \mathcal{D}$ may be a real number in the case of a point estimation problem. The best decision depends on unknown λ .

Assume that for $d \in \mathcal{D}$ there exists a loss function $L(\delta(x), \lambda) \geq 0$, the consequence of taking decision $\delta(x)$ when the distribution of X is $F_\lambda(x)$ and where δ is a function which assigns a decision $\delta(x) \in \mathcal{D}$ to each possible value x of the random variable X . For any δ , the expected loss when λ is the parameter is

$$\begin{aligned}
 R(\delta, \lambda) &= \int_{\mathcal{X}} L(\delta(x), \lambda) \, dF_{\lambda}(x) \\
 (2.1.1) \qquad &= \int_{\mathcal{X}} L(\delta(x), \lambda) \, f_{\lambda}(x) \, d\mu(x)
 \end{aligned}$$

since we can assume without loss of generality that $F_{\lambda}(x)$ is given in terms of its probability density $f_{\lambda}(x)$ with respect to some measure μ on the sample space. The overall expected loss (global risk) when the a priori distribution of Λ is G is

$$\begin{aligned}
 R(\delta, G) &= \int_{\Omega} R(\delta, \lambda) \, dG(\lambda) \\
 (2.1.2) \qquad &= \int_{\Omega} \int_{\mathcal{X}} L(\delta(x), \lambda) \, f_{\lambda}(x) \, d\mu(x) \, dG(\lambda) \\
 &= \int_{\mathcal{X}} \varphi_G(\delta(x), x) \, d\mu(x)
 \end{aligned}$$

where

$$(2.1.3) \qquad \varphi_G(d, x) = \int_{\Omega} L(d, \lambda) \, f_{\lambda}(x) \, dG(\lambda) \geq 0 \quad .$$

$R(\delta, G)$ is called the Bayes risk relative to G . If there exists a decision function δ_G such that for a.e. (μ) x

$$(2.1.4) \qquad \varphi_G(\delta_G(x), x) = \min_d \varphi_G(d, x) \quad ,$$

then since $\varphi_G \geq 0$, we have for any decision function δ ,

$$(2.1.5) \quad R(\delta_G, G) = \int_{\mathcal{X}} \min_d \varphi_G(d, x) d\mu(x) \leq R(\delta, G) .$$

If we define

$$(2.1.6) \quad \begin{aligned} R(G) &= R(\delta_G, G) = \int_{\mathcal{X}} \varphi_G(\delta_G(x), x) d\mu(x) \\ &= \min_{\delta} R(\delta, G) , \end{aligned}$$

then any decision function δ_G satisfying (2.1.4) minimizes $R(\delta, G)$ and is called a Bayes decision function corresponding to G , and $R(G)$ defined by (2.1.6) is called the Bayes envelope function. Thus assuming G exists, we can use $R(\delta, G)$ to judge how good any decision function δ is. Since G is unknown, δ_G is not directly available to us.

In the parametric empirical Bayes approach, assuming G is unknown, we suppose that the decision problem just described occurs repeatedly with independent random vector (Λ, X) . Thus we have the sequence

$$(2.1.7) \quad (\Lambda_1, X_1), (\Lambda_2, X_2), \dots, (\Lambda_n, X_n)$$

of independent pairs of random variables where the Λ_i 's, $i = 1, 2, \dots, n$, are identically distributed according to G , and where for $n = 1, 2, \dots$, the conditional distribution of X_n given that $\Lambda_n = \lambda$ is specified by the probability density $f_{\lambda}(x)$. When a decision about $\lambda_{n+1} = \lambda$

has to be made, we will have observed $x_1, x_2, \dots, x_n, x_{n+1} = x$ although the values $\lambda_1, \lambda_2, \dots$ remain always unknown. Thus for the decision about λ_{n+1} we can use a function of x_{n+1} whose form depends on $\underline{x}_n = x_1, x_2, \dots, x_n$ with \underline{X}_n the corresponding random variable; i.e. a function

$$(2.1.8) \quad \delta_n(\cdot) = \delta_n(\underline{X}_n; \cdot) \quad .$$

Then we take action $\delta_n(x) \in \mathcal{D}$ with loss $L(\delta_n(x), \lambda)$. Although the values x_1, x_2, \dots, x_n are independent of λ , these observations do contain information about G , the common unconditional density of the x_1, x_2, \dots, x_n, x with respect to μ being given by

$$(2.1.9) \quad f_G(x) = \int_{\Omega} f_{\lambda}(x) d G(\lambda) \quad .$$

It is hoped that for large n , δ_n will be close to the optimal but unknown δ_G which we would use throughout if G were known.

We define a sequential decision procedure to be the sequence $D = \{\delta_n\}$ of the form (2.1.8) for (2.1.7) with values in \mathcal{D} . From (2.1.2), the expected loss on the decision δ_n will be

$$(2.1.10) \quad \int_{\mathcal{X}} \varphi_G(\delta_n(x), x) d\mu(x) \quad ,$$

and hence the overall average loss will be

$$(2.1.11) \quad R_n(D, G) = \int_{\mathcal{X}} E \varphi_G(\delta_n(x), x) d\mu(x) \quad ,$$

where E denotes expectation with respect to the n independent random variables X_1, X_2, \dots, X_n . Since $\varphi_G(\delta_G(x), x) = \min_{\delta} \varphi_G(\delta(x), x)$, therefore $\varphi_G(\delta_G(x), x) \leq E \varphi_G(\delta_n(x), x)$, and hence

$$\begin{aligned} R(G) &= \int_{\mathcal{X}} \varphi_G(\delta_G(x), x) d\mu(x) \\ &\leq R_n(D; G) = \int_{\mathcal{X}} E \varphi_G(\delta_n(x), x) d\mu(x) . \end{aligned}$$

We wish to find a sequence of functions D which for every G contained in some class \mathcal{G} of a priori distributions is such that

$$(2.1.12) \quad \lim_{n \rightarrow \infty} R_n(D, G) = R(G) ,$$

the sequence of functions then being said to be asymptotically optimal.

2.2. General Results on Asymptotic Optimality.

Define

$$(2.2.1) \quad \Delta_G(d, x) = \int_{\Omega} [L(d, \lambda) - L(d_0, \lambda)] f_{\lambda}(x) d G(\lambda)$$

where d_0 is an arbitrary fixed element of \mathcal{D} . The following theorem is due to Robbins [14].

Theorem 1. Let G be such that

$$(2.2.2) \quad \int_{\Omega} L(\lambda) d G(\lambda) < \infty$$

where

$$(2.2.3) \quad 0 \leq L(\lambda) = \sup_d L(d, \lambda) \leq \infty .$$

Let

$$(2.2.4) \quad \Delta_n(d, x) = \Delta_n(X_1, X_2, \dots, X_n; d; x)$$

be a sequence of functions such that for a.e. $(\mu)x$,

$$(2.2.5) \quad \sup_d |\Delta_n(d, x) - \Delta_G(d, x)| \xrightarrow{P} 0 .$$

$$(2.2.6) \quad \text{If } \delta_n(x) = \delta_n(X_1, X_2, \dots, X_n; x) \text{ is any element } \bar{d} \in \mathcal{D} \text{ such} \\ \text{that } \Delta_n(\bar{d}, x) \leq \inf_d \Delta_n(d, x) + \epsilon_n$$

where $\epsilon_n \rightarrow 0$ as $n \rightarrow \infty$ is any sequence of constants, then $D = \{\delta_n\}$
is asymptotically optimal relative to G .

Proof. To prove that D is asymptotically optimal as defined by (2.1.12), it is sufficient, on account of the dominated convergence theorem, to show that

$$(2.2.7) \quad E \varphi_G(\delta_n(x), x) \leq H(x) \quad \text{for all } n \text{ where}$$

$$\int_{\mathcal{X}} H(x) d\mu(x) < \infty ,$$

and

$$(2.2.8) \quad \lim_{n \rightarrow \infty} E \varphi_G(\delta_n(x), x) = \varphi_G(\delta_G(x), x) \quad \text{a.e. } (\mu)x .$$

Letting

$$(2.2.9) \quad H(x) = \int_{\Omega} L(\lambda) f_{\lambda}(x) d G(\lambda) \geq 0 ,$$

we have from (2.1.3) and (2.2.3) that for any $D = \{\delta_n\}$,

$$(2.2.10) \quad \begin{aligned} \varphi_G(\delta_n(x), x) &= \int_{\Omega} L(\delta_n(x), \lambda) f_{\lambda}(x) d G(\lambda) \\ &\leq \int_{\Omega} L(\lambda) f_{\lambda}(x) d G(\lambda) = H(x) \quad \text{for all } n . \end{aligned}$$

Also,

$$(2.2.11) \quad \begin{aligned} \int_{\mathfrak{X}} H(x) d\mu(x) &= \int_{\Omega} L(\lambda) \left(\int_{\mathfrak{X}} f_{\lambda}(x) d\mu(x) \right) d G(\lambda) \\ &= \int_{\Omega} L(\lambda) d G(\lambda) < \infty \end{aligned}$$

by (2.2.2), so that from (2.2.10) and (2.2.11) we obtain (2.2.7). To prove (2.2.8) it will suffice to prove that

$$(2.2.12) \quad \varphi_G(\delta_n(x), x) \xrightarrow{P} \varphi_G(\delta_G(x), x) \quad \text{a.e. } (\mu)_x$$

since from (2.2.10) and (2.2.11) $\varphi_G(\delta_n(x), x) \leq H(x) < \infty$ a.e. $(\mu)_x$, enabling us to apply the dominated convergence theorem again. If

$$(2.2.13) \quad L_0(x) = \int_{\Omega} L(d_0, \lambda) f_{\lambda}(x) d G(\lambda) < \infty ,$$

மேலும்

$$\int_{\mathbb{R}^n} \nabla \cdot (f(x) \nabla u(x)) dx = 0 \quad (3.2)$$

இதன்

பொருள் $\int_{\mathbb{R}^n} \nabla \cdot (f(x) \nabla u(x)) dx = 0$ எனில் $\int_{\mathbb{R}^n} \nabla \cdot (f(x) \nabla u(x)) dx = 0$ எனில் $\int_{\mathbb{R}^n} \nabla \cdot (f(x) \nabla u(x)) dx = 0$

$$\int_{\mathbb{R}^n} \nabla \cdot (f(x) \nabla u(x)) dx = 0 \quad (3.3)$$

இதன் பொருள் $\int_{\mathbb{R}^n} \nabla \cdot (f(x) \nabla u(x)) dx = 0$

மேலும்

$$\int_{\mathbb{R}^n} \nabla \cdot (f(x) \nabla u(x)) dx = 0 \quad (3.4)$$

$$\int_{\mathbb{R}^n} \nabla \cdot (f(x) \nabla u(x)) dx = 0$$

இதன் பொருள் $\int_{\mathbb{R}^n} \nabla \cdot (f(x) \nabla u(x)) dx = 0$ எனில் $\int_{\mathbb{R}^n} \nabla \cdot (f(x) \nabla u(x)) dx = 0$ எனில் $\int_{\mathbb{R}^n} \nabla \cdot (f(x) \nabla u(x)) dx = 0$

$$\int_{\mathbb{R}^n} \nabla \cdot (f(x) \nabla u(x)) dx = 0 \quad (3.5)$$

இதன் பொருள் $\int_{\mathbb{R}^n} \nabla \cdot (f(x) \nabla u(x)) dx = 0$ எனில் $\int_{\mathbb{R}^n} \nabla \cdot (f(x) \nabla u(x)) dx = 0$ எனில் $\int_{\mathbb{R}^n} \nabla \cdot (f(x) \nabla u(x)) dx = 0$

$$\int_{\mathbb{R}^n} \nabla \cdot (f(x) \nabla u(x)) dx = 0 \quad (3.6)$$

then for a.e. $(\mu)x$,

$$(2.2.14) \quad \varphi_G(d, x) = L_0(x) + \Delta_G(d, x) \quad .$$

From (2.1.4) and (2.2.14) we have

$$\begin{aligned} \Delta_G(\delta_n(x), x) &= \varphi_G(\delta_n(x), x) - L_0(x) \\ &\geq \varphi_G(\delta_G(x), x) - L_0(x) = \Delta_G(\delta_G(x), x) \quad . \end{aligned}$$

Thus for any given $\epsilon > 0$ and n sufficiently large we have by (2.2.5) and (2.2.6) that, with probability as close to one as we please,

$$\begin{aligned} (2.2.15) \quad 0 &\leq \Delta_G(\delta_n(x), x) - \Delta_G(\delta_G(x), x) \\ &= [\Delta_G(\delta_n(x), x) - \Delta_n(\delta_n(x), x)] \\ &\quad + [\Delta_n(\delta_n(x), x) - \Delta_n(\delta_G(x), x)] \\ &\quad + [\Delta_n(\delta_G(x), x) - \Delta_G(\delta_G(x), x)] \\ &\leq \epsilon + \epsilon_n + \epsilon \quad . \end{aligned}$$

Therefore

$$(2.2.16) \quad \Delta_G(\delta_n(x), x) \xrightarrow{P} \Delta_G(\delta_G(x), x) \quad \text{a.e. } (\mu)x$$

and by (2.2.14),

$$\varphi_G(\delta_n(x), x) \xrightarrow{P} \varphi_G(\delta_G(x), x) \quad \text{a.e. } (\mu)x \quad ,$$

which proves (2.2.12) and hence the theorem.

When the decision space \mathcal{D} is finite, we have the following corollary, the proof of which is the same as for Theorem 1 with appropriate changes.

Corollary 1. Let $\mathcal{D} = \{d_0, d_1, \dots, d_m\}$ be a finite set and G be such that

$$(2.2.17) \quad \int_{\Omega} L(d_i, \lambda) dG(\lambda) < \infty, \quad i = 0, 1, \dots, m.$$

Let $\Delta_{i,n}(x) = \Delta_{i,n}(X_1, X_2, \dots, X_n; x)$ for $i = 1, 2, \dots, m$, and
 $n = 1, 2, \dots$, be such that for a.e. $(\mu)x$,

$$(2.2.18) \quad \Delta_{i,n}(x) \xrightarrow{P} \int_{\Omega} [L(d_i, \lambda) - L(d_0, \lambda)] f_{\lambda}(x) dG(\lambda) = \Delta_G(d_i, x).$$

Setting $\Delta_{0,n}(x) = 0$ and defining

$$(2.2.19) \quad \delta_n(x) = d_k \text{ where } k \text{ is any integer } 0 \leq k \leq m \text{ such}$$

$$\text{that } \Delta_{k,n}(x) = \min [\Delta_{0,n}(x) = 0, \Delta_{1,n}(x), \dots, \Delta_{m,n}(x)],$$

then $D = \{\delta_n\}$ is asymptotically optimal relative to G .

If $m = 1$, we have the two decision problem and obtain

Corollary 2. Let $\mathcal{D} = \{d_0, d_1\}$ and let G be such that

$$(2.2.20) \quad \int_{\Omega} L(d_i, \lambda) dG(\lambda) < \infty, \quad i = 0, 1,$$

and let $\Delta_n(x) = \Delta_n(X_1, X_2, \dots, X_n; x)$ be such that for a.e. $(\mu)x$,

$$(2.2.21) \quad \Delta_n(x) \xrightarrow{P} \Delta_G(x) = \int_{\Omega} [L(d_1, \lambda) - L(d_0, \lambda)] f_{\lambda}(x) d G(\lambda) .$$

Defining

$$(2.2.22) \quad \delta_n(x) = \begin{cases} d_0, & \text{if } \Delta_n(x) \geq 0 \\ d_1, & \text{if } \Delta_n(x) < 0 \end{cases} ,$$

then $D = \{\delta_n\}$ is asymptotically optimal relative to G .

2.3. Application for Two Decision Problems.

A. Discrete Case.

Samuel [20] and Robbins [13] and [14] have applied the above to the problem of hypothesis testing. Suppose we restrict ourselves to discrete distributions of the type

$$(2.3.1) \quad f_{\lambda}(x) = \lambda^x h(\lambda) g(x) \quad \text{for } x = 0, 1, 2, \dots .$$

Some of the more common distributions which follow (2.3.1) are the Poisson, geometric, and negative binomial distributions. μ will be counting measure on $\mathcal{X} = \{0, 1, 2, \dots\}$. Let us consider the case where

$$(2.3.2) \quad L(d_1, \lambda) - L(d_0, \lambda) = \sum_{j=0}^s a_j \lambda^j ,$$

that is, the difference between the losses is a polynomial in λ . For

(2.3.2), (2.2.20) holds whenever G has a finite moment of order s .

Then for (2.3.2) one has

$$(2.3.3) \quad \Delta_G(x) = \int_{\Omega} [L(d_1, \lambda) - L(d_0, \lambda)] f_{\lambda}(x) d G(\lambda)$$

$$= \int_{\Omega} \left(\sum_{j=0}^s a_j \lambda^j \right) f_{\lambda}(x) d G(\lambda)$$

and for distributions of the type (2.3.1), we have from (2.1.9),

$$(2.3.4) \quad f_G(x) = \int_{\Omega} g(x) \lambda^x h(\lambda) d G(\lambda) .$$

Thus in this case (2.3.3) becomes

$$\Delta_G(x) = \int_{\Omega} \left(\sum_{j=0}^s a_j \lambda^j \right) \lambda^x h(\lambda) g(x) d G(\lambda)$$

$$(2.3.5) \quad = \sum_{j=0}^s a_j g(x) \int_{\Omega} \lambda^{x+j} h(\lambda) d G(\lambda)$$

$$= \sum_{j=0}^s a_j f_G(x+j) \frac{g(x)}{g(x+j)} .$$

Now defining

$$(2.3.6) \quad \delta(x, y) = \begin{cases} 1 , & \text{if } x = y \\ 0 , & \text{if } x \neq y \end{cases}$$

and

$$(2.3.7) \quad f_n(x) = f_n(X_1, X_2, \dots, X_n; x) = \frac{1}{n} \sum_{j=1}^n \delta(X_j, x)$$

and since from (2.3.6)

$$(2.3.8) \quad E \delta(X_j, x) = P(X_j = x) = f_G(x) ,$$

it follows from the law of large numbers that

$$(2.3.9) \quad f_n(x) \xrightarrow{P} f_G(x) , \quad x = 0, 1, 2, \dots ,$$

and hence from (2.3.5) that

$$(2.3.10) \quad \Delta_n(x) = \sum_{j=0}^s a_j f_n(x+j) \frac{g(x)}{g(x+j)} \xrightarrow{P} \Delta_G(x) .$$

Thus for the hypothesis testing problem with loss function defined by (2.3.2) and for distributions of type (2.3.1), condition (2.2.21) is satisfied so that Corollary 2 is applicable and (2.2.22) yields an "optimal" empirical Bayes rule provided (2.2.20) holds.

Two common loss functions are particular cases of (2.3.2).

$$(2.3.11) \quad \begin{aligned} L(d_0, \lambda) &= \begin{cases} 0 & , \text{ if } \lambda \leq \lambda^* \\ \lambda - \lambda^* & , \text{ if } \lambda > \lambda^* \end{cases} \\ L(d_1, \lambda) &= \begin{cases} \lambda^* - \lambda & , \text{ if } \lambda \leq \lambda^* \\ 0 & , \text{ if } \lambda > \lambda^* \end{cases} \end{aligned}$$

where λ^* is a fixed constant, seems to be appropriate for the problem of testing a one-sided null hypothesis

$$(2.3.12) \quad H_0 : \lambda \leq \lambda^*$$

concerning the value of a parameter λ . The two actions available are $d_0 = \text{accept } H_0$ and $d_1 = \text{reject } H_0$. For testing two-sided hypotheses of the kind

$$(2.3.13) \quad H_0 : |\lambda - \lambda^*| \leq \Delta$$

where λ^* and $\Delta > 0$ are fixed constants,

$$L(d_0, \lambda) = \begin{cases} (\lambda - \lambda^*)^2 - \Delta^2, & \text{if } |\lambda - \lambda^*| > \Delta \\ 0, & \text{if } |\lambda - \lambda^*| \leq \Delta \end{cases}$$

$$(2.3.14)$$

$$L(d_1, \lambda) = \begin{cases} 0, & \text{if } |\lambda - \lambda^*| > \Delta \\ \Delta^2 - (\lambda - \lambda^*)^2, & \text{if } |\lambda - \lambda^*| \leq \Delta \end{cases}$$

seems to be appropriate. The two possible actions again are $d_0 = \text{accept } H_0$ and $d_1 = \text{reject } H_0$.

As a particular example, consider the problem of testing the one-sided null hypothesis given by (2.3.12) concerning the value of a Poisson parameter λ . Then $\Omega = \{0 < \lambda < \infty\}$ and we adopt the loss structure (2.3.11) with appropriate actions d_0 and d_1 . Then

$$f_\lambda(x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \dots,$$

and from (2.3.1) we see that $g(x) = \frac{1}{x!}$. Now

$$L(d_1, \lambda) - L(d_0, \lambda) = \lambda^* - \lambda = \sum_{j=0}^1 a_j \lambda^j$$

from (2.3.11) so that $a_0 = \lambda^*$ and $a_1 = -1$. Thus from (2.3.10) we get

$$\Delta_n(x) = \sum_{j=0}^1 a_j f_n(x+j) \frac{(x+j)!}{x!} = \lambda^* f_n(x) - (x+1) f_n(x+1).$$

Thus if

$$(2.3.15) \quad \delta_n(x) = \begin{cases} d_0, & \text{if } \frac{(x+1) f_n(x+1)}{f_n(x)} \leq \lambda^*, \\ d_1, & \text{if } \frac{(x+1) f_n(x+1)}{f_n(x)} > \lambda^*, \end{cases}$$

then $D = \{\delta_n\}$ is asymptotically optimal relative to every G such that

$$\int_0^\infty \lambda dG(\lambda) < \infty.$$

B. Continuous Case.

Corollary 2 may also be applied to the case where the independent and identically distributed random variables $\{X_i : i = 1, 2, \dots, n\}$ have a distribution function which is absolutely continuous with respect to Lebesgue measure μ on $\mathcal{X} = (-\infty, \infty)$.

Let $f_\lambda(x)$ be the conditional density of X given $\Lambda = \lambda$.

If Λ has the a priori distribution G on Ω , the (marginal) density

function of x will be

$$(2.3.16) \quad f_G(x) = \int_{\Omega} f_{\lambda}(x) d G(\lambda) .$$

In the empirical Bayesian situation, independent random variables $\{X_i : i = 1, 2, \dots, n\}$ constitute a random sample of size n from the distribution with density (2.3.16), and thus we wish to find an estimate $f_n(x) = f_n(\underline{X}_n; x)$ such that for all x and as $n \rightarrow \infty$

$$(2.3.17) \quad f_n(x) \xrightarrow{P} f_G(x)$$

for all possible f_G , $f_n(x)$ then being a consistent estimate of a density function.

One of the simplest classes of estimates satisfying (2.3.17) with some optimality properties and given in [15] is

$$(2.3.18) \quad f_n(x) = \frac{F_n(x+h_n) - F_n(x-h_n)}{2h_n}$$

where $h_n = dn^{-1/5}$, $d > 0$ is some constant, and where $F_n(x)$ is the empirical distribution function, viz.

$$(2.3.19) \quad F_n(x) = F_n(\underline{X}_n; x) = \frac{1}{n}(\text{number of indices } i, i = 1, 2, \dots, n \text{ for which } X_i \leq x) .$$

For the general case with loss functions satisfying (2.3.2) and density functions corresponding to (2.3.1) of the form

$$(2.3.20) \quad f_{\lambda}(x) = \begin{cases} \lambda^x g(x) h(\lambda) , & \text{for } a < x < \infty \\ 0 & , \text{ otherwise} \end{cases}$$

where a is some constant and may be $-\infty$, we have

$$(2.3.21) \quad \Delta_G(x) = \int_{\Omega} \left(\sum_{j=0}^s a_j \lambda^j \right) f_{\lambda}(x) dG(\lambda) .$$

It follows as before that when (2.3.17) holds,

$$(2.3.22) \quad \Delta_n(x) = \Delta_n(\underline{X}_n; x) = \sum_{j=0}^s a_j f_n(x+j) \frac{g(x)}{g(x+j)} \xrightarrow{P} \Delta_G(x) .$$

Thus if (2.2.20) holds, an optimal empirical Bayes rule for the problem of testing hypotheses with loss function defined by (2.3.2) is given by (2.2.22).

Many well-known density functions can be described by (2.3.20) after the application of a simple transformation. As a particular example, consider the gamma distribution given by

$$(2.3.23) \quad f_{\theta}(x) = \begin{cases} [(2\theta^2)^{p/2} \Gamma(\frac{p}{2})]^{-1} x^{(p/2)-1} \exp(-\frac{x}{2\theta^2}), & \text{if } x > 0 \\ 0 & , \text{ otherwise} \end{cases}$$

where $p > 0$ and $0 < \theta < \infty$. If we let $\lambda = \exp(-\frac{1}{2\theta^2})$, then (2.3.23) becomes

$$(2.3.24) \quad f_{\lambda}(x) = \begin{cases} \lambda^x x^{(p/2)-1} \frac{(-\log \lambda)^{p/2}}{\Gamma(\frac{p}{2})} , & \text{if } x > 0 \\ 0 & , \text{ otherwise} \end{cases}$$

where $0 < \lambda < 1$. This has the form of (2.3.20).

Since the function $\lambda = \psi(\theta)$ which takes the parameter θ (for the range of θ for which $f_{\theta}(x)$ is a density function) into λ is a strictly monotone function of θ , there is a one-one correspondence

between hypothesis (2.3.12) or (2.3.13) stated in terms of θ and the one stated in terms of λ . If the original loss function which is given as a function of θ can be written in the form (2.3.2) as a function of λ and if the a priori distribution $G^*(\theta)$ corresponds to the a priori distribution $G(\lambda)$, then the empirical Bayes rule given by (2.2.22) with $\Delta_n(x)$ given by (2.3.22) will be "optimal in the limit."

2.4. General Methods for Obtaining Consistent Estimators in the Two Decision Problem.

From (2.3.3) it is seen that

$$\begin{aligned} (2.4.1) \quad \Delta_G(x) &= \int_{\Omega} (\lambda^* - \lambda) f_{\lambda}(x) d G(\lambda) \\ &= \lambda^* f_G(x) - \int_{\Omega} \lambda f_{\lambda}(x) d G(\lambda) \end{aligned}$$

using loss function (2.3.11), and

$$\begin{aligned} (2.4.2) \quad \Delta_G(x) &= \int_{\Omega} [\Delta^2 - (\lambda - \lambda^*)^2] f_{\lambda}(x) d G(\lambda) \\ &= (\Delta^2 - \lambda^{*2}) f_G(x) - \int_{\Omega} \lambda^2 f_{\lambda}(x) d G(\lambda) + 2\lambda^* \int_{\Omega} \lambda f_{\lambda}(x) d G(\lambda) \end{aligned}$$

using loss function (2.3.14). Thus if $\int_{\Omega} \lambda d G(\lambda)$ and $\int_{\Omega} \lambda^2 d G(\lambda)$ are finite,

$$(2.4.3) \quad E(\Lambda | x) = \frac{\int_{\Omega} \lambda f_{\lambda}(x) d G(\lambda)}{f_G(x)},$$

$$(2.4.4) \quad E(\Lambda^2 | x) = \frac{\int_{\Omega} \lambda^2 f_{\lambda}(x) d G(\lambda)}{f_G(x)},$$

and if $E_n(\Lambda | x)$ and $E_n(\Lambda^2 | x)$ are such that

$$(2.4.5) \quad E_n(\Lambda | x) \xrightarrow{P} E(\Lambda | x)$$

$$(2.4.6) \quad E_n(\Lambda^2 | x) \xrightarrow{P} E(\Lambda^2 | x),$$

we have from Corollary 2 that equivalent sequential decision procedures to (2.2.22) for testing the hypotheses of the type (2.3.12) and (2.3.13), with loss functions (2.3.11) and (2.3.14) respectively, are given by

$$(2.4.7) \quad \delta_n(x) = \begin{cases} d_0, & \text{if } E_n(\Lambda | x) \leq \lambda^* \\ d_1, & \text{if } E_n(\Lambda | x) > \lambda^* \end{cases}$$

and

$$(2.4.8) \quad \delta_n(x) = \begin{cases} d_0, & \text{if } E_n(\Lambda^2 | x) - 2\lambda^* E_n(\Lambda | x) \leq \Delta^2 - \lambda^{*2} \\ d_1, & \text{if } E_n(\Lambda^2 | x) - 2\lambda^* E_n(\Lambda | x) > \Delta^2 - \lambda^{*2} \end{cases}.$$

Rutherford and Krutchkoff [16] give two general methods for obtaining the needed consistent estimators for $E(\Lambda | x)$ and $E(\Lambda^2 | x)$.

A. General Method I.

Define class \mathcal{F}_1 of families of conditional distributions by the following conditions:

- (a) In the family $\{F_\lambda(x); \lambda \in \Omega\}$, each $F_\lambda(x)$ has a density with respect to Lebesgue or counting measure.
- (b) The density functions $f_\lambda(x)$ are such that whenever $f_\lambda(x) > 0$,

$$(2.4.9) \quad \frac{1}{f_\lambda(x)} D_x^i f_\lambda(x) = Q_i(x, \lambda), \quad i = 0, 1, 2, \dots,$$

where $Q_i(x, \lambda)$ is a polynomial in λ of degree i and D_x^i represents the i 'th partial derivative if the measure is Lebesgue, or the i 'th finite difference if the measure is a counting measure.

For each member of any family in \mathcal{F}_1 one can always find functions $a_{ij}(x)$ such that

$$(2.4.10) \quad \lambda^j = \sum_{i=0}^j a_{ij}(x) Q_i(x, \lambda).$$

Hence we may write

$$\begin{aligned} E(\lambda^j | x) &= \frac{\int_{\Omega} \lambda^j f_\lambda(x) d G(\lambda)}{f_G(x)} \\ &= \int_{\Omega} \sum_{i=0}^j a_{ij}(x) Q_i(x, \lambda) \frac{f_\lambda(x)}{f_G(x)} d G(\lambda) \end{aligned}$$

$$\begin{aligned}
 (2.4.11) \quad &= \sum_{i=0}^j \frac{a_{ij}(x)}{f_G(x)} \int_{\Omega} Q_i(x, \lambda) f_{\lambda}(x) d G(\lambda) \\
 &= \sum_{i=0}^j \frac{a_{ij}(x)}{f_G(x)} \int_{\Omega} D_x^i f_{\lambda}(x) d G(\lambda) \\
 &= \sum_{i=0}^j \frac{a_{ij}(x)}{f_G(x)} D_x^i \int_{\Omega} f_{\lambda}(x) d G(\lambda) \\
 &= \sum_{i=0}^j a_{ij}(x) \frac{D_x^i f_G(x)}{f_G(x)}
 \end{aligned}$$

assuming that the order of differentiation and integration can be interchanged. Thus we require consistent estimators of $f_G(x)$, $D_x^1 f_G(x)$, and $D_x^2 f_G(x)$. Consistent estimators for $f_G(x)$ are given by (2.3.7) and (2.3.18) for the discrete and continuous cases respectively. Also, Rutherford has shown that

$$(2.4.12) \quad f_n^{[1]}(x) = \begin{cases} f_n(x+1) - f_n(x) & , \text{ for } X \text{ discrete} \\ \frac{f_n(x+h_n) - f_n(x)}{h_n} & , \text{ for } X \text{ continuous} \end{cases}$$

where $h_n = d n^{-1/5}$, $d > 0$ is some constant, and

$$(2.4.13) \quad f_n^{[2]}(x) = \begin{cases} f_n(x+2) - 2f_n(x+1) + f_n(x) & , \text{ for } X \text{ discrete} \\ \frac{f_n(x+2h_n) - 2f_n(x+h_n) + f_n(x)}{h_n^2} & , \text{ for } X \text{ continuous} \end{cases}$$

are consistent estimators for $D_x^1 f_G(x)$ and $D_x^2 f_G(x)$ respectively.

Thus

$$(2.4.14) \quad E_n(\Lambda^j | x) = \sum_{i=0}^j a_{ij}(x) \frac{f_n^{[i]}(x)}{f_n(x)}, \quad j = 1, 2,$$

are consistent estimators for $E(\Lambda | x)$ and $E(\Lambda^2 | x)$ respectively since the ratio of consistent estimators is a consistent estimator of the ratio.

As an example of this method consider the important case where X has a normal distribution with unknown mean λ and known variance σ^2 so that

$$(2.4.15) \quad f_\lambda(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{(x-\lambda)^2}{2\sigma^2} \right], \quad -\infty < x < \infty$$

where $\sigma > 0$ and $-\infty < \lambda < \infty$. The family of distributions corresponding to this density function is in the class \mathcal{F}_1 since condition (a) is satisfied and condition (b) is satisfied since for $i = 0$,

$$\frac{1}{f_\lambda(x)} D_x^0 f_\lambda(x) = 1 = Q_0(x, \lambda),$$

for $i = 1$,

$$\frac{\partial f_\lambda(x)}{\partial x} = \frac{1}{\sqrt{2\pi\sigma^2}} \frac{(x-\lambda)(-2)}{2\sigma^2} \exp \left[-\frac{(x-\lambda)^2}{2\sigma^2} \right] = \frac{(\lambda-x)}{\sigma^2} f_\lambda(x),$$

giving

$$\frac{1}{f_{\lambda}(x)} \frac{\partial f_{\lambda}(x)}{\partial x} = \frac{\lambda}{\sigma^2} - \frac{x}{\sigma^2} = Q_1(x, \lambda) ,$$

and for $i = 2$,

$$\frac{\partial^2 f_{\lambda}(x)}{\partial x^2} = -\frac{1}{\sigma^2} f_{\lambda}(x) + \frac{(\lambda-x)^2}{\sigma^4} f_{\lambda}(x) ,$$

giving

$$\frac{1}{f_{\lambda}(x)} \frac{\partial^2 f_{\lambda}(x)}{\partial x^2} = \frac{x^2}{\sigma^4} - \frac{2\lambda x}{\sigma^4} + \frac{\lambda^2}{\sigma^4} - \frac{1}{\sigma^2} = Q_2(x, \lambda) .$$

Then from (2.4.10) ,

$$\lambda = \sum_{i=0}^1 a_{i1}(x) Q_i(x, \lambda) = (x)(1) + (\sigma^2) \left(\frac{\lambda}{\sigma^2} - \frac{x}{\sigma^2} \right)$$

implies that $a_{01}(x) = x$ and $a_{11}(x) = \sigma^2$ and

$$\begin{aligned} \lambda^2 &= \sum_{i=0}^2 a_{i2}(x) Q_i(x, \lambda) \\ &= (x^2 + \sigma^2)(1) + (2x\sigma^2) \left(\frac{\lambda}{\sigma^2} - \frac{x}{\sigma^2} \right) + (\sigma^4) \left(\frac{x^2}{\sigma^4} - \frac{2\lambda x}{\sigma^4} + \frac{\lambda^2}{\sigma^4} - \frac{1}{\sigma^2} \right) \end{aligned}$$

implies that $a_{02} = x^2 + \sigma^2$, $a_{12} = 2x\sigma^2$, and $a_{22} = \sigma^4$. From

(2.4.14) we see that the consistent estimators for $E(\Lambda|x)$ and $E(\Lambda^2|x)$

are given by

$$(2.4.16) \quad E_n(\Lambda|x) = x + \frac{\sigma^2 f_n^{[1]}(x)}{f_n(x)}$$

and

$$(2.4.17) \quad E_n(\Lambda^2 | x) = x^2 + \sigma^2 + 2x\sigma^2 \frac{f_n^{[1]}(x)}{f_n(x)} + \sigma^4 \frac{f_n^{[2]}(x)}{f_n(x)}$$

respectively.

Other examples of the use of general Method I are:

(1) The gamma distribution with scale parameter λ and density

$$(2.4.18) \quad f_\lambda(x) = \begin{cases} \frac{1}{\Gamma(r)} \lambda^r x^{r-1} e^{-\lambda x} & , \quad 0 < x < \infty \\ 0 & , \quad \text{otherwise} \end{cases}$$

where $\lambda > 0$ and $r > 0$ are known. The consistent estimators are

$$(2.4.19) \quad E_n(\Lambda | x) = \frac{r-1}{x} - \frac{f_n^{[1]}(x)}{f_n(x)}$$

and

$$(2.4.20) \quad E_n(\Lambda^2 | x) = \frac{r(r-1)}{x^2} - \frac{2(r-1)}{x} \frac{f_n^{[1]}(x)}{f_n(x)} + \frac{f_n^{[2]}(x)}{f_n(x)} .$$

(2) The Poisson distribution with mean λ and density

$$(2.4.21) \quad f_\lambda(x) = \frac{e^{-\lambda} \lambda^x}{x!} , \quad x = 0, 1, 2, \dots ,$$

where $\lambda > 0$. The consistent estimators are

$$(2.4.22) \quad E_n(\Lambda | x) = (x+1) \left(\frac{f_n^{[1]}(x)}{f_n(x)} + 1 \right)$$

and

$$(2.4.23) \quad E_n(\Lambda^2 | x) = (x+1)(x+2) \left(\frac{f_n^{[2]}(x)}{f_n(x)} + 2 \frac{f_n^{[1]}(x)}{f_n(x)} + 1 \right)$$

(3) The geometric distribution with parameter λ and density

$$(2.4.24) \quad f_\lambda(x) = \lambda(1-\lambda)^{x-1}, \quad x = 1, 2, \dots,$$

where $0 \leq \lambda \leq 1$. The consistent estimators are

$$(2.4.25) \quad E_n(\Lambda | x) = - \frac{f_n^{[1]}(x)}{f_n(x)}$$

and

$$(2.4.26) \quad E_n(\Lambda^2 | x) = \frac{f_n^{[2]}(x)}{f_n(x)}.$$

(4) The negative binomial distribution with density

$$(2.4.27) \quad f_\lambda(x) = \binom{r+x-1}{x} (1-\lambda)^r \lambda^x, \quad x = 0, 1, 2, \dots,$$

where $0 \leq \lambda \leq 1$ and r is a fixed integer. The consistent estimators are

$$(2.4.28) \quad E_n(\Lambda | x) = \frac{x+1}{r+x} \left(\frac{f_n^{[1]}(x)}{f_n(x)} + 1 \right)$$

and

$$(2.4.29) \quad E_n(\Lambda^2 | x) = \frac{(x+1)(x+2)}{(r+x+1)(r+x)} \left[\frac{f_n^{[2]}(x)}{f_n(x)} + \frac{2f_n^{[1]}(x)}{f_n(x)} + 1 \right] .$$

(5) The logarithmic distribution with density

$$(2.4.30) \quad f_\lambda(x) = \frac{-1}{\log(1-\lambda)} \frac{\lambda^x}{x}, \quad x = 1, 2, \dots,$$

where $0 \leq \lambda < 1$. The consistent estimators are

$$(2.4.31) \quad E_n(\Lambda | x) = \frac{x+1}{x} \left[\frac{f_n^{[1]}(x)}{f_n(x)} + 1 \right]$$

and

$$(2.4.32) \quad E_n(\Lambda^2 | x) = \frac{x+2}{x} \left[\frac{f_n^{[2]}(x)}{f_n(x)} + \frac{2f_n^{[1]}(x)}{f_n(x)} + 1 \right]$$

B. General Method II.

Assume that X is a vector of $k \geq 3$ independent random variables each with the distribution function $F_\lambda(x)$. A family will be in the class \mathcal{F}_2 if

(c) there is a statistic T sufficient for λ , and

(d) the distribution of T has a density with respect to either Lebesgue or a counting measure of the form

$$(2.4.33) \quad f_\lambda(t, k) = \begin{cases} \varphi(k) \left(\frac{t}{\lambda}\right)^k h(t, \lambda), & \text{for all } t \text{ in some interval} \\ 0, & \text{otherwise} \end{cases}$$

where $\varphi(k)$ is a function of k only and $h(t, \lambda)$ is independent of k . Considering the first ν random variables in the vector only, the corresponding sufficient statistic T_ν has density

$$(2.4.34) \quad f_\lambda(t, \nu) = \begin{cases} \varphi(\nu) \left(\frac{t}{\lambda}\right)^\nu h(t, \lambda) & , \text{ for all } t \text{ in some interval} \\ 0 & , \text{ otherwise} \end{cases}$$

Then since

$$(2.4.35) \quad \begin{aligned} \lambda^j f_\lambda(t, k) &= \lambda^j \varphi(k) \left(\frac{t}{\lambda}\right)^k h(t, \lambda) \\ &= \frac{t^j \varphi(k)}{\varphi(k-j)} \varphi(k-j) \left(\frac{t}{\lambda}\right)^{k-j} h(t, \lambda) \\ &= \frac{t^j \varphi(k)}{\varphi(k-j)} f_\lambda(t, k-j) \end{aligned}$$

and if

$$(2.4.36) \quad \begin{aligned} \int_{\Omega} f_\lambda(t, k) d G(\lambda) &= f_G(t, k) \quad , \\ E(\Lambda^j | t) &= \int_{\Omega} \frac{\lambda^j f_\lambda(t, k)}{f_G(t, k)} d G(\lambda) \\ &= \frac{t^j \varphi(k)}{\varphi(k-j) f_G(t, k)} \int_{\Omega} f_\lambda(t, k-j) d G(\lambda) \\ &= \frac{t^j \varphi(k)}{\varphi(k-j)} \cdot \frac{f_G(t, k-j)}{f_G(t, k)} \quad . \end{aligned}$$

Letting $T_{\nu, i}$ represent the sufficient statistic based on

the first v recorded random variables in the i 'th replication of the problem, and defining from (2.3.7) and (2.3.18) ,

$$(2.4.37) \quad f_n(t, k-j) = \begin{cases} f_n(T_{k-j,1}, T_{k-j,2}, \dots, T_{k-j,n}; t) & , \quad \text{for } T \text{ discrete} , \\ \frac{F_n(T_{k-j,1}, T_{k-j,2}, \dots, T_{k-j,n}; t+h_n) - F_n(T_{k-j,1}, T_{k-j,2}, \dots, T_{k-j,n}; t-h_n)}{2h_n} & , \\ \end{cases} \quad \text{for } T \text{ continuous} ,$$

it follows that consistent estimators for $E(\Lambda|t)$ and $E(\Lambda^2|t)$ are

$$(2.4.38) \quad E_n(\Lambda^j|t) = \frac{t^j \varphi(k)}{\varphi(k-j)} \cdot \frac{f_n(t, k-j)}{f_n(t, k)} , \quad j = 1, 2 ,$$

respectively. For a sufficient statistic T , $E(\Lambda|t)$ and $E(\Lambda^2|t)$ are equivalent to $E(\Lambda|x)$ and $E(\Lambda^2|x)$ respectively.

As an example of general Method II, consider the case where X has an exponential distribution with parameter $\frac{1}{\lambda}$. Assume that we have observations on k independent and identically distributed random variables, each having the conditional density function

$$(2.4.39) \quad f_\lambda(x) = \begin{cases} \frac{1}{\lambda} e^{-\frac{x}{\lambda}} & , \quad 0 < x < \infty \\ 0 & , \quad \text{otherwise} \end{cases}$$

where $\lambda > 0$.

A statistic T will be sufficient for λ if

$$(2.4.40) \quad L(x_1, x_2, \dots, x_v | \lambda) = g(t | \lambda) k(x_1, x_2, \dots, x_v)$$

where $L(x_1, x_2, \dots, x_v | \lambda)$ is the likelihood for the sample values x_1, x_2, \dots, x_v given λ , $g(t | \lambda)$ is the conditional density function for T , and $k(x_1, x_2, \dots, x_v)$ is independent of λ . Considering only the first v independent observations, let

$$t = \sum_{i=1}^v x_i .$$

We now derive the conditional density function for T . Now

$$f_{\lambda}(x_1, x_2, \dots, x_v) = \prod_{i=1}^v f_{\lambda}(x_i) = \frac{1}{\lambda^v} \exp \left[-\frac{1}{\lambda} \left(\sum_{i=1}^v x_i \right) \right]$$

where $\min(x_1, x_2, \dots, x_v) > 0$. If we use the transformation

$$u_1 = x_1 + x_2 + \dots + x_v, \quad u_2 = x_2, \dots, \quad u_v = x_v,$$

then

$$x_1 = u_1 - u_2 - \dots - u_v, \quad x_2 = u_2, \dots, \quad x_v = u_v,$$

and the Jacobian of the transformation is

$$\frac{\partial(x_1, x_2, \dots, x_v)}{\partial(u_1, u_2, \dots, u_v)} = \begin{vmatrix} \frac{\partial x_1}{\partial u_1} & \frac{\partial x_1}{\partial u_2} & \dots & \frac{\partial x_1}{\partial u_v} \\ \frac{\partial x_2}{\partial u_1} & \frac{\partial x_2}{\partial u_2} & \dots & \frac{\partial x_2}{\partial u_v} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \frac{\partial x_v}{\partial u_1} & \frac{\partial x_v}{\partial u_2} & \dots & \frac{\partial x_v}{\partial u_v} \end{vmatrix} = \begin{vmatrix} 1 & -1 & -1 & \dots & -1 & -1 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ 0 & 0 & 0 & \dots & 0 & 1 \end{vmatrix} = 1$$

Then

$$f_\lambda(u_1, u_2, \dots, u_v) = f_\lambda(x_1(u_1, \dots, u_v), x_2(u_1, \dots, u_v), \dots, x_v(u_1, \dots, u_v)) \left| \frac{\partial(x_1, x_2, \dots, x_v)}{\partial(u_1, u_2, \dots, u_v)} \right|$$

$$= \frac{1}{\lambda^v} e^{-\frac{1}{\lambda} u_1}, \quad \text{if } u_1 - u_2 - \dots - u_v > 0, \quad u_2 > 0, \dots, u_v > 0,$$

and

$$f_\lambda(u_1) = \int_0^{u_1 - u_2 - \dots - u_{v-1}} \int_0^{u_1 - u_2 - \dots - u_{v-2}} \dots \int_0^{u_1} \frac{1}{\lambda^v} e^{-\frac{u_1}{\lambda}} du_2 du_3 \dots du_v$$

$$= \begin{cases} \frac{1}{\lambda^v} \frac{1}{(v-1)!} u_1^{v-1} e^{-\frac{u_1}{\lambda}}, & \text{if } u_1 > 0 \\ 0, & \text{if } u_1 \leq 0 \end{cases}.$$

Thus

$$\begin{aligned}
 L(x_1, x_2, \dots, x_\nu | \lambda) &= \frac{1}{\lambda^\nu} e^{-\frac{1}{\lambda} \left(\sum_{i=1}^{\nu} x_i \right)} = \frac{1}{\lambda^\nu} e^{-\frac{u_1}{\lambda}} \\
 &= \left(\frac{1}{\Gamma(\nu)} \frac{u_1^{\nu-1}}{\lambda^\nu} e^{-\frac{u_1}{\lambda}} \right) \left(\frac{\Gamma(\nu)}{u_1^{\nu-1}} \right) \\
 &= f_\lambda(u_1) k(x_1, x_2, \dots, x_\nu)
 \end{aligned}$$

and by (2.4.40), u_1 is sufficient for λ since

$$k(x_1, x_2, \dots, x_\nu) = \frac{\Gamma(\nu)}{\left(\sum_{i=1}^{\nu} x_i \right)^{\nu-1}}$$

is independent of λ . Therefore

$$T = \sum_{i=1}^{\nu} x_i$$

is a sufficient statistic for λ and has conditional density function

$$(2.4.41) \quad f_\lambda(t, \nu) = \begin{cases} \frac{1}{\lambda \Gamma(\nu)} \left(\frac{t}{\lambda} \right)^\nu \left(\frac{\lambda}{t} \right) e^{-\frac{t}{\lambda}}, & 0 < t < \infty \\ 0, & \text{otherwise} \end{cases}$$

Comparing (2.4.41) and (2.4.34) we see that $\varphi(k) = \frac{1}{\lambda \Gamma(k)}$ so that from (2.4.38) the consistent estimates of $E(\Lambda | x)$ and $E(\Lambda^2 | x)$ are

$$(2.4.42) \quad E_n(\Lambda | x) = E_n(\Lambda | t) = \frac{\frac{t}{\lambda \Gamma(k)}}{\frac{1}{\lambda \Gamma(k-1)}} \frac{f_n(t, k-1)}{f_n(t, k)} = \frac{t}{k-1} \frac{f_n(t, k-1)}{f_n(t, k)}$$

and

$$(2.4.43) \quad E_n(\Lambda^2 | x) = E_n(\Lambda^2 | t) = \frac{t^2}{(k-1)(k-2)} \frac{f_n(t, k-2)}{f_n(t, k)}$$

respectively.

Other examples of the use of general Method II are:

(1) The normal distribution with known mean μ , unknown variance λ , and with conditional density function

$$(2.4.44) \quad f_{\lambda}(x) = \frac{1}{\sqrt{2\pi\lambda}} \exp \left[-\frac{(x-\mu)^2}{2\lambda} \right], \quad -\infty < x < \infty$$

where $\lambda > 0$. If we consider the first 2ν of the observations on $2k$ independent and identically distributed random variables each with conditional density function (2.4.44), then

$$T = \sum_{i=1}^{2\nu} (X_i - \mu)^2$$

is a sufficient statistic for λ with conditional density

$$(2.4.45) \quad f_{\lambda}(t, \nu) = \begin{cases} \frac{1}{2^{\nu} \Gamma(\nu)} \left(\frac{t}{\lambda}\right)^{\nu} e^{-\frac{t}{2\lambda}}, & 0 < t < \infty. \\ 0 & \text{otherwise.} \end{cases}$$

The consistent estimators are

$$(2.4.46) \quad E_n(\Lambda | x) = \frac{t}{2k-2} \frac{f_n(t, k-1)}{f_n(t, k)}$$

and

$$(2.4.47) \quad E_n(\Lambda^2 | x) = \frac{t^2}{(2k-2)(2k-4)} \frac{f_n(t, k-2)}{f_n(t, k)}.$$

(2) The uniform distribution with one known terminal point, range λ , and with conditional density function

$$(2.4.48) \quad f_{\lambda}(x) = \begin{cases} \frac{1}{\lambda}, & 0 < x \leq \lambda < \infty. \\ 0 & \text{otherwise.} \end{cases}$$

If we consider the first ν of the observations on k independent and identically distributed random variables each with conditional density function (2.4.48), then $T = \max [X_1, X_2, \dots, X_\nu]$ is a sufficient statistic for λ with conditional density given by

$$(2.4.49) \quad f_\lambda(t, \nu) = \begin{cases} \frac{\nu}{t} \left(\frac{t}{\lambda}\right)^\nu, & 0 < t \leq \lambda < \infty \\ 0, & \text{otherwise} \end{cases}$$

The consistent estimators are

$$(2.4.50) \quad E_n(\Lambda | x) = \frac{tk}{k-1} \frac{f_n(t, k-1)}{f_n(t, k)}$$

and

$$(2.4.51) \quad E_n(\Lambda^2 | x) = \frac{t^2 k}{k-2} \frac{f_n(t, k-2)}{f_n(t, k)}.$$

2.5. Point Estimation.

Using the loss function (2.3.11), suppose that instead of testing the hypothesis (2.3.12) we wish to estimate the unknown parameter λ . If $\mathcal{D} = \{d\}$ is the set of allowed estimates of λ , we have proved in Theorem 1 that if $\delta(x)$ is any estimator of λ and if

$$(2.5.1) \quad \int_{\Omega} L(\lambda) dG(\lambda) < \infty$$

and

$$(2.5.2) \quad \varphi_G(\delta_n(\underline{X}_n; x), x) \xrightarrow{P} \varphi_G(\delta(x), x) \quad \text{a.e. } (\mu)_x,$$

then

$$\lim_{n \rightarrow \infty} E L(\delta_n(\underline{X}_n; X), \Lambda) = E L(\delta(X), \Lambda)$$

where $\delta(x)$ is a Bayes estimator. That is, the conditions (2.5.1) and

(2.5.2) imply that $\delta_n(\underline{X}_n; \mathbf{x})$, $n = 1, 2, \dots$, is an asymptotically optimal sequence of estimators.

Considering the special case of a loss function which is continuous in "d", Rutherford and Krutchkoff [16] give the following simplification of Robbins' result.

Lemma 1. If

(2.5.3) $L(d, \lambda)$ is continuous in d for each $\lambda \in \Omega$,

(2.5.4) $\int_{\Omega} L(\lambda) \, dG(\lambda) < \infty$,

and

(2.5.5) $\delta_n(\underline{X}_n; \mathbf{x}) \xrightarrow{P} \delta(\mathbf{x}) \quad \text{a.e. } (\mu)\mathbf{x}$,

where $\delta(\mathbf{x})$ is a Bayes estimator, then $\delta_n(\underline{X}_n; \mathbf{x})$, $n = 1, 2, \dots$, is an asymptotically optimal sequence of estimators.

Proof. From (2.5.4),

$$\begin{aligned} \int_{\Omega} L(\lambda) \, dG(\lambda) &= \int_{\Omega} L(\lambda) \left[\int_{\mathbf{x}} f_{\lambda}(\mathbf{x}) \, d\mu(\mathbf{x}) \right] dG(\lambda) \\ &= \int_{\Omega} \int_{\mathbf{x}} L(\lambda) f_{\lambda}(\mathbf{x}) \, d\mu(\mathbf{x}) \, dG(\lambda) < \infty \end{aligned}$$

implies that $L(\lambda) f_{\lambda}(\mathbf{x})$ is integrable. Since $L(d, \lambda) f_{\lambda}(\mathbf{x})$ is bounded uniformly by $L(\lambda) f_{\lambda}(\mathbf{x})$, then using a result of Cramér (p. 57 of [1]),

if $L(d, \lambda)$ is continuous in d for each λ , then

$$\varphi_G(d, \mathbf{x}) = \int_{\Omega} L(d, \lambda) f_{\lambda}(\mathbf{x}) d G(\lambda)$$

is continuous in d . Then (2.5.5) implies (2.5.2) and therefore

$\delta_n(\underline{X}_n; \mathbf{x})$ is an asymptotically optimal sequence of estimators.

For the usual squared error loss function which is continuous, let $\omega(\mathbf{x})$ be an estimate of λ . The loss function when \mathbf{x} is observed is

$$(2.5.6) \quad (\omega(\mathbf{x}) - \lambda)^2.$$

Then, in the discrete case, the expected squared deviation is

$$\begin{aligned} & E(\omega(X) - \Lambda)^2 \\ &= E\{E[(\omega(X) - \Lambda)^2 | \Lambda = \lambda]\} \\ (2.5.7) \quad &= E\left\{\sum_{\mathbf{x}} f_{\lambda}(\mathbf{x}) (\omega(\mathbf{x}) - \lambda)^2\right\} \\ &= \int_{\Omega} \sum_{\mathbf{x}} f_{\lambda}(\mathbf{x}) (\omega(\mathbf{x}) - \lambda)^2 d G(\lambda) \\ &= \text{minimum} \end{aligned}$$

when we define $\omega(\mathbf{x})$ for each value of \mathbf{x} as that value $y = y(\mathbf{x})$ for which

$$\begin{aligned} (2.5.8) \quad I(\mathbf{x}) &= \int_{\Omega} f_{\lambda}(\mathbf{x}) (y - \lambda)^2 d G(\lambda) \\ &= \text{minimum} . \end{aligned}$$

For any fixed x ,

$$\begin{aligned}
 I(x) &= \int_{\Omega} f_{\lambda}(x) (y^2 - 2y\lambda + \lambda^2) dG(\lambda) \\
 &= y^2 \int_{\Omega} f_{\lambda}(x) dG(\lambda) - 2y \int_{\Omega} \lambda f_{\lambda}(x) dG(\lambda) \\
 (2.5.9) \quad &+ \int_{\Omega} \lambda^2 f_{\lambda}(x) dG(\lambda) \\
 &= \int_{\Omega} f_{\lambda}(x) dG(\lambda) \left[y - \frac{\int_{\Omega} \lambda f_{\lambda}(x) dG(\lambda)}{\int_{\Omega} f_{\lambda}(x) dG(\lambda)} \right]^2 \\
 &+ \int_{\Omega} \lambda^2 f_{\lambda}(x) dG(\lambda) - \frac{\left(\int_{\Omega} \lambda f_{\lambda}(x) dG(\lambda) \right)^2}{\int_{\Omega} f_{\lambda}(x) dG(\lambda)}
 \end{aligned}$$

and therefore

$$\frac{d}{dy} I(x) = 2 \int_{\Omega} f_{\lambda}(x) dG(\lambda) \left[y - \frac{\int_{\Omega} \lambda f_{\lambda}(x) dG(\lambda)}{\int_{\Omega} f_{\lambda}(x) dG(\lambda)} \right] = 0$$

when

$$y = \frac{\int_{\Omega} \lambda f_{\lambda}(x) dG(\lambda)}{\int_{\Omega} f_{\lambda}(x) dG(\lambda)}$$

and $I(x)$ will be a minimum. Thus the Bayes estimator of λ with respect to an a priori distribution G of λ is

$$(2.5.10) \quad \omega_G(x) = \frac{\int_{\Omega} \lambda f_{\lambda}(x) d G(\lambda)}{\int_{\Omega} f_{\lambda}(x) d G(\lambda)} = E(\Lambda|x) .$$

It follows from Lemma 1 that if

$$\int_{\Omega} L(\lambda) d G(\lambda) < \infty ,$$

then any consistent sequence of estimators $E_n(\Lambda|x)$ for $E(\Lambda|x)$ is asymptotically optimal. Such estimators can then be determined by general Methods I and II in section 2.4 for a wide class of distributions.

When the set of allowed estimates $\mathcal{D} = \{d\}$ is bounded, then

$$(2.5.11) \quad L(\lambda) = \sup_d (d - \lambda)^2$$

is bounded for each $\lambda \in \Omega$. This implies that

$$\int_{\Omega} L(\lambda) d G(\lambda) < \infty$$

and hence a sufficient condition to ensure that $E_n(\Lambda|x)$, $n = 1, 2, \dots$, is an asymptotically optimal sequence of estimators is that

$$(2.5.12) \quad E_n(\Lambda|x) \xrightarrow{P} E(\Lambda|x) \quad \text{a.e. } (\mu)^x .$$

If the set of allowed estimates $\mathcal{D} = \{d\}$ is unbounded,

$$\int_{\Omega} L(\lambda) d G(\lambda)$$

is not finite and therefore the result is not true. However when Λ has a bounded fourth moment in its a priori distribution G , it will be shown that a consistent sequence of estimators for $E(\Lambda|x)$ will allow us to find an asymptotically optimal sequence of estimators in a squared error loss, point estimation problem.

Let

$$(2.5.13) \quad E(\Lambda^4) = \int_{\Omega} \lambda^4 d G(\lambda) \leq B < \infty$$

and define for any $\epsilon > 0$ and each x ,

$$(2.5.14) \quad E^{B, \epsilon}(\Lambda|x) = \begin{cases} \frac{B^{3/4}}{\epsilon} & , \text{ when } E(\Lambda|x) > \frac{B^{3/4}}{\epsilon} \\ E(\Lambda|x) & , \text{ when } \frac{-B^{3/4}}{\epsilon} \leq E(\Lambda|x) \leq \frac{B^{3/4}}{\epsilon} \\ \frac{-B^{3/4}}{\epsilon} & , \text{ when } E(\Lambda|x) < \frac{-B^{3/4}}{\epsilon} \end{cases}$$

and

$$(2.5.15) \quad E_n^{B, \epsilon}(\Lambda|x) = \begin{cases} \frac{B^{3/4}}{\epsilon} & , \text{ when } E_n(\Lambda|x) > \frac{B^{3/4}}{\epsilon} \\ E_n(\Lambda|x) & , \text{ when } \frac{-B^{3/4}}{\epsilon} \leq E_n(\Lambda|x) \leq \frac{B^{3/4}}{\epsilon} \\ \frac{-B^{3/4}}{\epsilon} & , \text{ when } E_n(\Lambda|x) < \frac{-B^{3/4}}{\epsilon} \end{cases} .$$

We have then

Theorem 2. If

$$(2.5.16) \quad E(\Lambda^4) \leq B < \infty$$

and

$$(2.5.17) \quad E_n(\Lambda|x) \xrightarrow{P} E(\Lambda|x) \quad \text{a.e.}(\mu)_x, ,$$

then for any $\epsilon > 0$,

$$(2.5.18) \quad \lim_{n \rightarrow \infty} E[E_n^{B, \epsilon}(\Lambda|X) - \Lambda]^2 \leq R + \epsilon$$

where

$$R = E[E(\Lambda|X) - \Lambda]^2 .$$

Proof. From (2.5.17), we see that

$$(2.5.19) \quad E_n^{B, \epsilon}(\Lambda|x) \xrightarrow{P} E^{B, \epsilon}(\Lambda|x)$$

for any $\epsilon > 0$. Suppose that our set of allowed estimates $\mathcal{D} = \{d\}$ is truncated at $-B$ and B . This gives the bounded case and, using Lemma 1, condition (2.5.19) ensures that $E_n^{B, \epsilon}(\Lambda|x)$, $n = 1, 2, \dots$, is an asymptotically optimal sequence of estimators. Therefore

$$(2.5.20) \quad \begin{aligned} \lim_{n \rightarrow \infty} E[E_n^{B, \epsilon}(\Lambda|X) - \Lambda]^2 \\ = E[E^{B, \epsilon}(\Lambda|X) - \Lambda]^2 . \end{aligned}$$

But

$$\begin{aligned}
 & E[E^B, \epsilon(\Lambda|X) - \Lambda]^2 \\
 &= E[E^B, \epsilon(\Lambda|X) - E(\Lambda|X) + E(\Lambda|X) - \Lambda]^2 \\
 (2.5.21) \quad &= E[E^B, \epsilon(\Lambda|X) - E(\Lambda|X)]^2 \\
 &+ 2E[(E^B, \epsilon(\Lambda|X) - E(\Lambda|X))(E(\Lambda|X) - \Lambda)] \\
 &+ E[E(\Lambda|X) - \Lambda]^2.
 \end{aligned}$$

The second term of (2.5.21) is zero since

$$\begin{aligned}
 & E[(E^B, \epsilon(\Lambda|X) - E(\Lambda|X))(E(\Lambda|X) - \Lambda)] \\
 &= E\{E[(E^B, \epsilon(\Lambda|X) - E(\Lambda|X))(E(\Lambda|X) - \Lambda) | X]\} \\
 (2.5.22) \quad &= E\{[E^B, \epsilon(\Lambda|X) - E(\Lambda|X)][E(\Lambda|X) - E(\Lambda|X)]\} \\
 &= 0,
 \end{aligned}$$

and the first term of (2.5.21) becomes

$$\begin{aligned}
 & E[E^B, \epsilon(\Lambda|X) - E(\Lambda|X)]^2 \\
 (2.5.23) \quad &= \int_{\mathcal{X}} [E^B, \epsilon(\Lambda|x) - E(\Lambda|x)]^2 d F_G(x) \\
 &\leq \int_{\beta} [E(\Lambda|x)]^2 d F_G(x)
 \end{aligned}$$

where

$$\beta = \{x : |E(\Lambda|x)| > \frac{B^{3/4}}{\epsilon}\}$$

and $F_G(x)$ is the marginal distribution of X . Applying the Schwarz and Hölder inequalities successively to (2.5.23) we get

$$\begin{aligned}
 & E[E^{B,\epsilon}(\Lambda|X) - E(\Lambda|X)]^2 \\
 (2.5.24) \quad & \leq \sqrt{\int_{\beta} [E(\Lambda|x)]^4 dF_G(x) \int_{\beta} dF_G(x)} \\
 & \leq \sqrt{\int_{\beta} E(\Lambda^4|x) dF_G(x) \int_{\beta} dF_G(x)} .
 \end{aligned}$$

Now by (2.5.16) ,

$$\begin{aligned}
 \int_{\beta} E(\Lambda^4|x) dF_G(x) & \leq \int_{\mathcal{X}} E(\Lambda^4|x) dF_G(x) \\
 (2.5.25) \quad & = E(\Lambda^4) \\
 & \leq B < \infty
 \end{aligned}$$

and

$$\begin{aligned}
 (2.5.26) \quad \int_{\beta} dF_G(x) & = P[|E(\Lambda|X)| > \frac{B^{3/4}}{\epsilon}] \\
 & \leq \frac{\text{Var} [E(\Lambda|X)] \epsilon^2}{B^{3/2}}
 \end{aligned}$$

by Tchebychev's inequality. Also, using the Schwarz inequality, we have

$$(2.5.27) \quad \text{Var} [E(\Lambda|X)] \leq \text{Var} (\Lambda) \leq E(\Lambda^2) \leq \sqrt{E(\Lambda^4)} = \sqrt{B}$$

and so

$$(2.5.28) \quad \int_{\beta} dF_G(x) \leq \frac{\epsilon^2}{B} .$$

Substituting (2.5.25) and (2.5.28) in (2.5.24), we see that

$$(2.5.29) \quad E[E^{B, \epsilon}(\Lambda|X) - E(\Lambda|X)]^2 \leq \epsilon.$$

Therefore we see from (2.5.21) that

$$(2.5.30) \quad \lim_{n \rightarrow \infty} E[E_n^{B, \epsilon}(\Lambda|X) - \Lambda]^2 \leq R + \epsilon.$$

CHAPTER III

ESTIMATING THE PRIOR DISTRIBUTION

3.1. Robbins General Method.

Returning to Corollary 2, recall that for $\mathcal{D} = \{d_0, d_1\}$ and $\Omega = \{-\infty < \lambda < \infty\}$, an asymptotically optimal decision procedure D exists relative to the class

$$\mathcal{G} = \{G : \int_{\Omega} L(d_i, \lambda) dG(\lambda) < \infty, \quad i = 0, 1\}$$

whenever we can find a sequence $\Delta_n(x) = \Delta_n(X_1, X_2, \dots, X_n; x)$ such that for a.e. $(\mu)x$,

$$(3.1.1) \quad \Delta_n(x) \xrightarrow{P} \Delta_G(x) = \int_{\Omega} [L(d_1, \lambda) - L(d_0, \lambda)] f_{\lambda}(x) dG(\lambda)$$

for every $G \in \mathcal{G}$. We wish to construct a sequence $\Delta_n(x)$ by finding a sequence $G_n(\lambda) = G_n(X_1, X_2, \dots, X_n; \lambda)$ of random distribution functions in λ such that

$$(3.1.2) \quad P[\lim_{n \rightarrow \infty} G_n(\lambda) = G(\lambda) \text{ at every continuity point } \lambda \text{ of } G] = 1.$$

Setting

$$(3.1.3) \quad \Delta_n(x) = \int_{-\infty}^{\infty} [L(d_1, \lambda) - L(d_0, \lambda)] f_{\lambda}(x) dG_n(\lambda),$$

and if the function

$$(3.1.4) \quad [L(d_1, \lambda) - L(d_0, \lambda)] f_\lambda(x)$$

is continuous in λ for a.e. (μ) fixed x , then by the Helly-Bray theorem (p. 180 of [9]), (3.1.1) holds.

Robbins [14] gives the following method for constructing a particular sequence $G_n(\lambda)$ of random estimators of unknown $G(\lambda)$. Assume that for every $\lambda \in \Omega = \{-\infty < \lambda < \infty\}$, $F_\lambda(x)$ is a specified distribution function in x , and for every fixed $x \in \mathcal{X} = \{-\infty < x < \infty\}$, $F_\lambda(x)$ is a Borel measurable function of λ . Define for any distribution G of Λ a distribution function in x given by

$$(3.1.5) \quad F_G(x) = \int_{-\infty}^{\infty} F_\lambda(x) dG(\lambda) .$$

Letting X_1, X_2, \dots , be a sequence of independent random variables with common distribution function F_G , define

$$(3.1.6) \quad B_n(x) = B_n(X_1, X_2, \dots, X_n; x) = \frac{1}{n} (\text{number of terms } X_i, \\ i = 1, 2, \dots, n \text{ which are } \leq x) .$$

The distance between any two distribution functions F_1, F_2 is defined to be

$$(3.1.7) \quad \rho(F_1, F_2) = \sup_x |F_1(x) - F_2(x)| .$$

Let \mathcal{G} be any class of distribution functions of Λ such that $G \in \mathcal{G}$.

If $G_n(\lambda) = G_n(X_1, X_2, \dots, X_n; \lambda) \in \mathcal{G}$ and is such that

$$(3.1.8) \quad \rho(B_n, F_{G_n}) \leq \frac{1}{G} \inf_{G \in \mathcal{G}} \rho(B_n, F_{\frac{1}{G}}) + \epsilon_n$$

where ϵ_n is any sequence of constants tending to zero as $n \rightarrow \infty$, then the sequence G_n is said to be effective for \mathcal{G} if (3.1.2) holds for every $G \in \mathcal{G}$.

Robbins then proves the following theorem which ensures that under suitable conditions on the family of densities $f_\lambda(x)$, the relation (3.1.2) holds whatever G is.

Theorem 3. Assume that

$$(3.1.9) \quad \text{for every fixed } x, \quad F_\lambda(x) \text{ is a continuous function of } \lambda,$$

$$(3.1.10) \quad \text{the limits } F_{-\infty}(x) = \lim_{\lambda \rightarrow -\infty} F_\lambda(x) \text{ and } F_\infty(x) = \lim_{\lambda \rightarrow \infty} F_\lambda(x) \\ \text{exist for every } x,$$

$$(3.1.11) \quad \text{neither } F_{-\infty} \text{ nor } F_\infty \text{ is a distribution function,}$$

and

$$(3.1.12) \quad \text{if } G_1, G_2 \text{ are any two distribution functions in } \lambda \\ \text{such that } F_{G_1} = F_{G_2}, \text{ then } G_1 = G_2.$$

Then the sequence G_n defined by (3.1.8) is effective for the class \mathcal{G}

of all distribution functions in λ .

Proof: By the Glivenko-Cantelli theorem (p. 20 of [9]), we have that

$$(3.1.13) \quad P\left[\lim_{n \rightarrow \infty} \rho(B_n, F_G) = 0\right] = 1.$$

Now using (3.1.8),

$$(3.1.14) \quad \begin{aligned} \rho(F_{G_n}, F_G) &\leq \rho(F_{G_n}, B_n) + \rho(B_n, F_G) \\ &\leq \inf_{G \in \mathcal{Y}} \rho(B_n, F_G) + \epsilon_n + \rho(B_n, F_G) \\ &\leq \rho(B_n, F_G) + \epsilon_n + \rho(B_n, F_G). \end{aligned}$$

It follows from (3.1.13) that with probability one, $\rho(F_{G_n}, F_G) \rightarrow 0$;
i.e. with probability one the sequence X_1, X_2, \dots , is such that

$$(3.1.15) \quad \lim_{n \rightarrow \infty} \int_{-\infty}^{\infty} F_{\lambda}(x) dG_n(\lambda) = \int_{-\infty}^{\infty} F_{\lambda}(x) dG(\lambda)$$

uniformly in x .

Consider any fixed sequence X_1, X_2, \dots , such that (3.1.15) holds. Let G_{k_n} be any subsequence of G_n such that $G_{k_n}(\lambda) \rightarrow G^*(\lambda)$ at every continuity point λ of G^* , G^* being a "weak" distribution function ($G^*(-\infty) \geq 0$, $G^*(\infty) \leq 1$). Using the Helly-Bray theorem, we have from (3.1.9) and (3.1.10) that for every x ,

$$(3.1.16) \quad \lim_{n \rightarrow \infty} \int_{-\infty}^{\infty} F_{\lambda}(x) dG_{k_n}(\lambda) = \int_{-\infty}^{\infty} F_{\lambda}(x) dG^*(\lambda) + G^*(-\infty)F_{-\infty}(x) + [1 - G^*(\infty)]F_{\infty}(x)$$

of the following conditions (1.1) - (1.4):

(1.1) $\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n f_k(x) = f(x)$ a.e. on X .

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n f_k(x) = f(x) \quad (1.1.1)$$

where $f(x) \in L^1(X)$.

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n f_k(x) = f(x) \quad (1.1.2)$$

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n f_k(x) = f(x) \quad (1.1.3)$$

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n f_k(x) = f(x) \quad (1.1.4)$$

where $f(x) \in L^1(X)$ and $f_k(x) \in L^1(X)$ for all k . The conditions (1.1) - (1.4) are satisfied if $f(x) \in L^1(X)$ and $f_k(x) \in L^1(X)$ for all k .

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n f_k(x) = f(x) \quad (1.1.5)$$

where $f(x) \in L^1(X)$.

where $f(x) \in L^1(X)$ and $f_k(x) \in L^1(X)$ for all k . The conditions (1.1) - (1.4) are satisfied if $f(x) \in L^1(X)$ and $f_k(x) \in L^1(X)$ for all k . The conditions (1.1) - (1.4) are satisfied if $f(x) \in L^1(X)$ and $f_k(x) \in L^1(X)$ for all k . The conditions (1.1) - (1.4) are satisfied if $f(x) \in L^1(X)$ and $f_k(x) \in L^1(X)$ for all k .

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n f_k(x) = f(x) \quad (1.1.6)$$

and hence from (3.1.15),

$$(3.1.17) \quad \int_{-\infty}^{\infty} F_{\lambda}(x) dG(\lambda) = \int_{-\infty}^{\infty} F_{\lambda}(x) dG^{*}(\lambda) + G^{*}(-\infty) F_{-\infty}(x) + [1 - G^{*}(\infty)] F_{\infty}(x)$$

for every x .

Let us show now that G^{*} is a distribution function; i.e. that $G^{*}(-\infty) = 0$ and $G^{*}(\infty) = 1$. $F_{-\infty}$ is the limit as $\lambda \rightarrow -\infty$ of F_{λ} and hence is a nondecreasing function of x . By (3.1.11),

$$(3.1.18) \quad 0 \leq F_{-\infty}(-\infty) \quad \text{and} \quad F_{-\infty}(\infty) \leq 1.$$

Similarly, F_{∞} is a nondecreasing function of x and

$$(3.1.19) \quad 0 \leq F_{\infty}(-\infty), \quad F_{\infty}(\infty) \leq 1.$$

If we let $x \rightarrow -\infty$, then by Lebesgue's bounded convergence theorem and (3.1.17),

$$(3.1.20) \quad G^{*}(-\infty) F_{-\infty}(-\infty) + [1 - G^{*}(\infty)] F_{\infty}(-\infty) = 0$$

and hence if $G^{*}(-\infty) \neq 0$, then $F_{-\infty}(-\infty) = 0$, and if $G^{*}(\infty) \neq 1$, then $F_{\infty}(-\infty) = 0$. Similarly if $x \rightarrow \infty$ in (3.1.17), then

$$(3.1.21) \quad G^{*}(-\infty) F_{-\infty}(\infty) + [1 - G^{*}(\infty)] F_{\infty}(\infty) = 1 - G^{*}(\infty) + G^{*}(-\infty)$$

and hence if $G^{*}(-\infty) \neq 0$, then $F_{-\infty}(\infty) = 1$, and if $G^{*}(\infty) \neq 1$, then $F_{\infty}(\infty) = 1$.

If a_n is any sequence of constants converging to a limit a from the right, then subtracting (3.1.17) with $x = a$ from (3.1.17) with $x = a_n$ where $n \rightarrow \infty$, we have that

$$(3.1.22) \quad G^*(-\infty)[F_{-\infty}(a+0) - F_{-\infty}(a)] + [1 - G^*(\infty)][F_{\infty}(a+0) - F_{\infty}(a)] = 0.$$

Therefore $G^*(-\infty) \neq 0$ implies that $F_{-\infty}(a+0) = F_{-\infty}(a)$, and $G^*(\infty) \neq 1$ implies that $F_{\infty}(a+0) = F_{\infty}(a)$. Since a distribution function F is by definition nondecreasing, continuous on the right, and such that

$$\lim_{x \rightarrow -\infty} F(x) = 0 \quad \text{and} \quad \lim_{x \rightarrow \infty} F(x) = 1, \quad \text{it follows that if } G^*(-\infty) \neq 0$$

and $G^*(\infty) \neq 1$, then $F_{-\infty}$, F_{∞} are distribution functions which contradicts (3.1.11). Thus $G^*(-\infty) = 0$ and $G^*(\infty) = 1$; i.e. G^* is a distribution function. Now $F_G = F_{G^*}$ from (3.1.17), and therefore $G = G^*$ from (3.1.12). Since G^* denoted the weak limit of any convergent subsequence of G_n , (3.1.2) holds and the theorem is proved.

Theorem 3 can be appropriately modified when the parameter space Ω is not the whole line. If $\Omega = \{0 \leq \lambda < \infty\}$, for example, we obtain

Theorem 4. Assume that

$$(3.1.23) \quad \text{for every fixed } x, \quad F_{\lambda}(x) \text{ is a continuous function of } \lambda,$$

$$(3.1.24) \quad \text{the limit } F_{\infty}(x) = \lim_{\lambda \rightarrow \infty} F_{\lambda}(x) \text{ exists for every } x,$$

(3.1.25) F_{∞} is not a distribution function,

and

(3.1.26) if G_1, G_2 are any two distribution functions in λ which assign unit probability to $\Omega = \{0 \leq \lambda < \infty\}$ such that

$$\int_{\Omega} F_{\lambda}(x) d G_1(\lambda) = \int_{\Omega} F_{\lambda}(x) d G_2(\lambda) \text{ for all } x ,$$

then $G_1 = G_2$.

Then the sequence G_n defined by (3.1.8) is effective for the class
 \mathcal{J} of all distributions which assign unit probability to Ω .

As an example of Theorem 4 consider the case where we have a Poisson parameter. Thus for $\lambda = 0$, let

$$(3.1.27) \quad F_0(x) = \begin{cases} 0 , & \text{for } x < 0 \\ 1 , & \text{for } x \geq 0 \end{cases}$$

and for $0 < \lambda < \infty$, let

$$(3.1.28) \quad F_{\lambda}(x) = \sum_{0 \leq i \leq x} \frac{e^{-\lambda} \lambda^i}{i!} .$$

Conditions (3.1.23), (3.1.24), and (3.1.25) are satisfied, and it will be shown that (3.1.26) is also satisfied.

If $G \in \mathcal{J}$, then

$$(3.1.29) \quad F_G(x) = \int_{\Omega} F_{\lambda}(x) dG(\lambda) = \sum_{0 \leq i \leq x} \int_{\Omega} f_{\lambda}(i) dG(\lambda)$$

where

$$f_0(i) = \begin{cases} 1, & \text{for } i = 0, \\ 0, & \text{for } i = 1, 2, \dots, \end{cases}$$

(3.1.30)

$$f_{\lambda}(i) = \frac{e^{-\lambda} \lambda^i}{i!} \quad \text{for } i = 0, 1, \dots, \text{ and } 0 < \lambda < \infty.$$

$$\text{Now } F_G(0) = \int_{\Omega} f_{\lambda}(0) dG(\lambda), \quad F_G(n) - F_G(n-1) = \int_{\Omega} f_{\lambda}(n) dG(\lambda),$$

$n = 1, 2, \dots$, and $F_{G_1} = F_{G_2}$ implies that

$$(3.1.31) \quad \int_{\Omega} f_{\lambda}(n) dG_1(\lambda) = \int_{\Omega} f_{\lambda}(n) dG_2(\lambda), \quad n = 0, 1, 2, \dots$$

Define the set functions

$$(3.1.32) \quad H_j(B) = \frac{\int_B e^{-\lambda} dG_j(\lambda)}{\int_{\Omega} e^{-\lambda} dG_j(\lambda)}, \quad j = 1, 2,$$

where B is a Borel set in Ω . Then H_j is a probability measure on the Borel sets. Since from (3.1.31) and (3.1.30),

$$\begin{aligned} (3.1.33) \quad c &= \int_{\Omega} e^{-\lambda} dG_1(\lambda) = \int_{\Omega} f_{\lambda}(0) dG_1(\lambda) \\ &= \int_{\Omega} f_{\lambda}(0) dG_2(\lambda) = \int_{\Omega} e^{-\lambda} dG_2(\lambda), \end{aligned}$$

we can write

$$(3.1.34) \quad H_j(B) = \frac{1}{c} \int_B e^{-\lambda} d G_j(\lambda) , \quad j = 1, 2$$

and where $0 < c < \infty$. Then

$$(3.1.35) \quad \frac{d H_j}{d G_j} = \frac{e^{-\lambda}}{c} ,$$

and for $n = 1, 2, \dots$, and $j = 1, 2$,

$$(3.1.36) \quad \begin{aligned} \int_{\Omega} \lambda^n d H_j(\lambda) &= \frac{1}{c} \int_{\Omega} e^{-\lambda} \lambda^n d G_j(\lambda) \\ &= \frac{n!}{c} \int_{\Omega} f_{\lambda}(n) d G_j(\lambda) . \end{aligned}$$

Therefore we have from (3.1.31) that

$$(3.1.37) \quad \alpha_n = \int_{\Omega} \lambda^n d H_1(\lambda) = \int_{\Omega} \lambda^n d H_2(\lambda) ,$$

and $n = 1, 2, \dots$, $0 \leq f_{\lambda}(n) \leq 1$, imply that $0 \leq e^{-\lambda} \lambda^n \leq n!$ for $0 \leq \lambda < \infty$. Thus

$$(3.1.38) \quad 0 \leq \alpha_n = \frac{1}{c} \int_{\Omega} e^{-\lambda} \lambda^n d G_j(\lambda) \leq \frac{n!}{c}$$

for $n = 1, 2, \dots$, and hence

$$\sum_{n=1}^{\infty} \left(\frac{\alpha_n}{n!} \right) \left(\frac{1}{2} \right)^n \leq \frac{1}{c} \sum_{n=1}^{\infty} \left(\frac{1}{2} \right)^n < \infty .$$

From a theorem by Cramér (p. 176 of [1]) which says that if $\alpha_0 = 1, \alpha_1, \alpha_2, \dots$, are the moments of a certain distribution function $F(x)$, all of which are assumed to be finite, and the series

$$\sum_{r=0}^{\infty} \frac{\alpha_r}{r!} t^r$$

is absolutely convergent for some $t > 0$, then $F(x)$ is the only distribution function that has moments $\alpha_0, \alpha_1, \dots$, it follows from above that $H_1 = H_2$. Since

$$(3.1.39) \quad G_j(B) = \int_B \frac{d G_j}{d H_j} d H_j(\lambda) = \int_B c e^{\lambda} d H_j(\lambda), \quad j = 1, 2,$$

then $G_1(B) = G_2(B)$; i.e. $G_1 = G_2$ which shows that (3.1.26) holds and hence the sequence G_n defined by (3.1.8) is effective.

3.2. Squared Error Consistent Estimates of the Prior Distribution Function

Unlike the method of Robbins where the choice of G_n satisfying (3.1.8) is nonconstructive, Rutherford and Krutchkoff [17], [18] give the following method of using the sequence of observations x_1, x_2, \dots, x_n to give squared error consistent estimates of the prior distribution function $G(\lambda)$. Von Mises [25] in 1942 motivated the technique which will be developed by showing that if the first two moments of the prior distribution $G(\lambda)$ are known, then exact upper and lower bounds for $G(\lambda)$ can be found.

The general class \mathcal{G}_p which has a known subclass \mathcal{G} to which $G(\lambda)$ belongs, and the class \mathcal{F}_p to which the known conditional distribution function $F_\lambda(x)$ belongs, are described below. The class \mathcal{G}_p is defined by the following conditions:

- (a) $G(\lambda)$ is absolutely continuous with respect to Lebesgue measure.
- (b) Its density function $g(\lambda)$ is determined completely and continuously by its first $p \geq 2$ moments in some open p interval.

The subclass \mathcal{G} of \mathcal{G}_p is defined by the additional condition:

- (c) All $g(\lambda)$ in \mathcal{G} are determined by the same known continuous function described in (b).

Thus, for example, when the fourth moment exists, it can be shown that if the skewness and kurtosis are known, then the prior distribution, which will be a member of the Pearson family of curves, is determined completely. The Pearson family of curves can be represented by the solutions of the first order differential equation

$$(3.2.1) \quad \frac{dy}{dx} = \frac{y(m-x)}{a+bx+cx^2}$$

where the shapes of the curves depend on the parameters a, b, c , and m which will be known if the skewness and kurtosis are known. The class \mathcal{F}_p is defined by the condition

- (d) for each $F_\lambda(x) \in \mathcal{F}_p$ there exists known functions $h_k(\cdot)$, $k = 1, 2, \dots, p$, such that

$$(3.2.2) \quad E[h_k(X) | \lambda] = \lambda^k .$$

As an example of condition (d), suppose the observations were distributed according to a Poisson density with mean λ and that λ was distributed according to an unknown Pearson distribution. Then for $p = 4$ the functions $h_k(x)$ are given by $h_1(x) = x$, $h_2(x) = x(x-1)$, $h_3(x) = x(x-1)(x-2)$, and $h_4(x) = x(x-1)(x-2)(x-3)$. Thus, for example,

$$\begin{aligned} E[h_1(X) | \lambda] &= \sum_{x=0}^{\infty} x f_{\lambda}(x) = \sum_{x=1}^{\infty} \frac{\lambda^x e^{-\lambda}}{(x-1)!} \\ &= e^{-\lambda} \left(\lambda + \frac{\lambda^2}{1!} + \frac{\lambda^3}{2!} + \dots \right) \\ &= \lambda e^{-\lambda} e^{\lambda} = \lambda . \end{aligned}$$

Similarly we see that $E[h_k(X) | \lambda] = \lambda^k$, $k = 2, 3$, and 4 .

Taking expectations with respect to Λ in (3.2.2) we get

$$(3.2.3) \quad E(\Lambda^k) = E[E(h_k(X) | \Lambda)] = E[h_k(X)]$$

for $k = 1, 2, \dots, p$, and these are all finite because of condition (b).

Define the functions

$$(3.2.4) \quad M_k(\underline{x}_n) = \frac{1}{n} \sum_{i=1}^n h_k(x_i), \quad k = 1, 2, \dots, p$$

where \underline{x}_n represents the sequence of observations x_1, x_2, \dots, x_n with

corresponding random variable \underline{X}_n , and the vectors

$$(3.2.5) \quad \underline{M}(\underline{x}_n) = (M_1(\underline{x}_n), M_2(\underline{x}_n), \dots, M_p(\underline{x}_n))$$

and

$$(3.2.6) \quad \underline{\mu} = (E(\Lambda), E(\Lambda^2), \dots, E(\Lambda^p)) .$$

From (3.2.3) and (3.2.4) we see that

$$(3.2.7) \quad E[\underline{M}(\underline{X}_n)] = \underline{\mu}$$

and since $E(\Lambda^p) < \infty$, we have by the strong law of large numbers that

$$(3.2.8) \quad \underline{M}(\underline{X}_n) \xrightarrow{\text{a.s.}} \underline{\mu}$$

and thus $\underline{M}(\underline{X}_n)$ provides estimates of the first p moments of the prior distribution. Assume that the prior density functions $g(\lambda)$ belong to a specific subclass \mathcal{G} whose members we denote by $g(\lambda; \underline{\mu})$, the notation indicating the dependence of the $g(\lambda)$ on their moments.

The estimator of $g(\lambda; \underline{\mu})$ will be the density function $g(\lambda; \underline{M}(\underline{x}_n))$. For $p = 4$, $g(\lambda; \underline{M}(\underline{x}_n))$ represents the solution of Pearson's differential equation (3.2.1) with $\underline{M}(\underline{x}_n)$ substituted for $\underline{\mu}$. Since from condition (b), $g(\lambda)$ are continuous functions of $\underline{\mu}$ for every λ , and (3.2.8) holds,

$$(3.2.9) \quad g(\lambda; \underline{M}(\underline{X}_n)) \xrightarrow{\text{a.s.}} g(\lambda; \underline{\mu}) \quad \text{a.e. } \lambda .$$

We define our estimate of $G(\lambda)$ by

$$(3.2.10) \quad G_n(\lambda; \underline{M}(\underline{x}_n)) = \int_{-\infty}^{\lambda} g(t; \underline{M}(\underline{x}_n)) dt ,$$

and prove

Theorem 5. If $G \in \mathcal{G}$ and $F \in \mathcal{F}_p$, then

$$(3.2.11) \quad \lim_{n \rightarrow \infty} E[G_n(\lambda; \underline{M}(\underline{X}_n)) - G(\lambda)]^2 = 0 \quad \text{a.e. } \lambda$$

where the expectation is with respect to \underline{X}_n .

Proof. Defining

$$(3.2.12) \quad G_n^*(\lambda; \underline{M}(\underline{x}_n)) = \int_{-n}^{\lambda} g(t; \underline{M}(\underline{x}_n)) dt ,$$

and

$$(3.2.13) \quad G^*(\lambda) = \int_{-n}^{\lambda} g(t; \underline{\mu}) dt ,$$

we obtain

$$(3.2.14) \quad \begin{aligned} |G_n(\lambda; \underline{M}(\underline{x}_n)) - G(\lambda)| &\leq |G_n(\lambda; \underline{M}(\underline{x}_n)) - G_n^*(\lambda; \underline{M}(\underline{x}_n))| \\ &\quad + |G_n^*(\lambda; \underline{M}(\underline{x}_n)) - G^*(\lambda)| \\ &\quad + |G^*(\lambda) - G(\lambda)| . \end{aligned}$$

Each term on the right hand side of (3.2.14) converges almost surely to zero a.e. λ , which implies that

$$(3.2.15) \quad G_n(\lambda; \underline{M}(\underline{x}_n)) \xrightarrow{\text{a.s.}} G(\lambda), \quad \text{a.e. } \lambda.$$

Since $[G_n(\lambda; \underline{M}(\underline{x}_n)) - G(\lambda)]^2$ is bounded for all n , then from the dominated convergence theorem we obtain (3.2.11) and the proof is complete.

CHAPTER IV

SELECTING THE BEST OF K POPULATIONS

PARAMETRIC CASE

4.1. Introduction

Suppose we have k populations (categories, varieties, etc.) for which we can observe random variables X_i whose distribution depends on an unknown parameter λ_i for $i = 1, 2, \dots, k$. Define the "best" population as that population with the largest λ_i . The empirical Bayes approach has been used by Deely [2] to obtain an optimal decision procedure which will either (1) select the best population; or (2) select a subset of k populations which contains the best.

The observable random variable X_i has for each of the k populations the density $f_{\lambda_i}(x)$ with respect to some measure μ_i on the sample space where the unknown parameter $\lambda_i \in \Omega$, a subset of the real numbers. Define $\underline{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_k)$ to be the realization of the random vector $\underline{\Lambda} = (\Lambda_1, \Lambda_2, \dots, \Lambda_k)$ where $\underline{\lambda} \in \Omega^k$, a subset of Euclidean k -space. Assuming that the Λ_i 's, $i = 1, 2, \dots, k$ are independent random variables, we define G_i to be an a priori distribution function of Λ_i , and

$$(4.1.1) \quad G(\underline{\lambda}) = \prod_{j=1}^k G_j(\lambda_j)$$

to be an a priori distribution function on Ω^k . The observation $\underline{x} = (x_1, x_2, \dots, x_k)$ of the random vector $\underline{X} = (X_1, X_2, \dots, X_k)$ is in \mathcal{X}^k , a subset of Euclidean k -space.

If we wish to select one population and say it is the best, then an appropriate action space is $\mathcal{D} = \mathcal{D}_1 = \{d_1, d_2, \dots, d_k\}$ where action d_i means: "say λ_i is the largest." If we wish to select a subset of the k populations and say that the best is in this subset, then our action space is $\mathcal{D} = \mathcal{D}_2 = \{S_1, S_2, \dots, S_p\}$ where $p = 2^k - 1$ is the number of possible subsets of the integers $\{1, 2, \dots, k\}$ where the empty set is excluded. Action S_i means: "say subset S_i contains the best population."

Suppose that for $d_i \in \mathcal{D}_1$, $i = 1, 2, \dots, k$ or $S_i \in \mathcal{D}_2$, $j = 1, 2, \dots, p$ there exists a loss function $L(\delta(\underline{x}), \underline{\lambda}) \geq 0$, the consequence of taking a decision in either \mathcal{D}_1 or \mathcal{D}_2 when $\underline{\lambda}$ is the true parameter vector and where δ is a function which assigns a decision $\delta(\underline{x}) \in \mathcal{D}_1$ or $\delta(\underline{x}) \in \mathcal{D}_2$ to each possible value \underline{x} of the random vector \underline{X} . The expected loss when $\underline{\lambda}$ is the parameter vector is

$$(4.1.2) \quad R(\delta, \underline{\lambda}) = \int_{\mathcal{X}^k} L(\delta(\underline{x}), \underline{\lambda}) f_{\underline{\lambda}}(\underline{x}) d\mu(\underline{x}) ,$$

where $f_{\underline{\lambda}}(\underline{x}) = \prod_{i=1}^k f_{\lambda_i}(x_i)$ with respect to measure $\mu = \mu_1 \times \mu_2 \times \dots \times \mu_k$.

The overall expected loss when the a priori distribution of $\underline{\lambda}$ is $G(\underline{\lambda})$ given by (4.1.1) is

$$\begin{aligned}
 R(\delta, G) &= \int_{\Omega^k} R(\delta, \underline{\lambda}) d G(\underline{\lambda}) \\
 (4.1.3) \quad &= \int_{\Omega^k} \int_{\mathcal{X}^k} L(\delta(\underline{x}), \underline{\lambda}) f_{\underline{\lambda}}(\underline{x}) d\mu(\underline{x}) d G(\underline{\lambda}) \\
 &= \int_{\mathcal{X}^k} \phi_G(\delta, \underline{x}) d\mu(\underline{x}) ,
 \end{aligned}$$

where

$$(4.1.4) \quad \phi_G(\delta, \underline{x}) = \int_{\Omega^k} L(\delta(\underline{x}), \underline{\lambda}) f_{\underline{\lambda}}(\underline{x}) d G(\underline{\lambda})$$

by Fubini's theorem if $L(\delta(\underline{x}), \underline{\lambda}) f_{\underline{\lambda}}(\underline{x})$ is integrable on $\mathcal{X}^k \times \Omega^k$.

If there exists a decision function δ_G such that for a.e. $(\underline{\mu})_{\underline{x}}$,

$$(4.1.5) \quad \phi_G(\delta_G(\underline{x}), \underline{x}) \leq \phi_G(\delta(\underline{x}), \underline{x})$$

for every $\delta(\underline{x})$, we have

$$(4.1.6) \quad R(\delta_G, G) \leq R(\delta, G)$$

and δ_G is called a Bayes decision procedure with respect to G . For a finite action space $\mathcal{D} = \{d_1, d_2, \dots, d_m\}$ we have from (4.1.5) that δ_G is defined by

$$\begin{aligned}
 (4.1.7) \quad \delta_G(\underline{x}) &= d_j \text{ where } j \text{ is any integer } 1, 2, \dots, m \text{ such} \\
 \text{that } \phi_G(d_j, \underline{x}) &= \min_{1 \leq i \leq m} \{\phi_G(d_i, \underline{x})\} .
 \end{aligned}$$

δ_G is not directly available to us unless we have a complete knowledge of G . Assume that G exists but is unknown and that the decision problem occurs repeatedly giving us the random vectors

$$(4.1.8) \quad (\underline{x}_1, \underline{\Lambda}_1), (\underline{x}_2, \underline{\Lambda}_2), \dots, (\underline{x}_n, \underline{\Lambda}_n) ,$$

which are independent and such that the $\underline{\Lambda}_i$'s, $i = 1, 2, \dots, n$ are identically distributed according to G . The sequence (4.1.8) then contains information about the form of the unknown G .

If

$$(4.1.9) \quad \delta_n(\cdot) = \delta_n(\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n; \cdot)$$

is a mapping from $\mathcal{X}^{k(n+1)}$ to \mathcal{D} and takes action $\delta_n(\underline{x}) \in \mathcal{D}$ with loss $L(\delta_n(\underline{x}), \underline{\lambda})$, then for the given sequence $D = \{\delta_n\}$, the expected loss on the decision δ_n will be

$$(4.1.10) \quad \int_{\mathcal{X}^k} \phi_G(\delta_n(\underline{x}), \underline{x}) d\mu(\underline{x}) .$$

The overall average loss will be

$$(4.1.11) \quad R_n(D, G) = \int_{\mathcal{X}^k} E_n \phi_G(\delta_n(\underline{x}), \underline{x}) d\mu(\underline{x}) ,$$

where E_n denotes expectation with respect to the n independent random vectors $\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n$; i.e.

$$(4.1.12) \quad E_n \phi_G(\delta_n(\underline{x}), \underline{x}) = \int_{\mathcal{X}^k} \dots \int_{\mathcal{X}^k} \phi_G(\delta_n(\underline{x}), \underline{x}) f_G(\underline{x}) d\mu(\underline{x}) \dots f_G(\underline{x}) d\mu(\underline{x}) ,$$

where

$$(4.1.13) \quad f_G(\underline{x}) = \int_{\Omega^k} f_{\underline{\lambda}}(\underline{x}) d G(\underline{\lambda}) ,$$

and

$$f_{\underline{\lambda}}(\underline{x}) = \prod_{i=1}^k f_{\lambda_i}(x_i) .$$

The sequence D is said to be **asymptotically** optimal with respect to G if

$$(4.1.14) \quad \lim_{n \rightarrow \infty} R(\delta_n, G) = R(\delta_G, G)$$

for every $G \in \mathcal{G}$, some class of a priori distributions on Ω^k , and $\delta_n \in D$ is called an empirical Bayes procedure.

If r observations are taken from each of the k populations, then we get an $r \times k$ observation matrix $\underline{x}^* \in \mathcal{X}^{kr}$ with the corresponding random matrix \underline{X}^* . When $\underline{x}_1^*, \underline{x}_2^*, \dots, \underline{x}_n^*$ are the prior observations of the random matrix \underline{X}^* , we can define an empirical Bayes procedure in an analogous manner.

4.2. Selecting a Subset Containing the Best Population.

It will now be shown that the Bayes procedure for selecting the best out of k populations is precisely the same as the Bayes procedure which selects a subset of k populations which contains the

best for a particular loss function.

Considering $\mathcal{A}_2 = \{S_1, S_2, \dots, S_p\}$ where $p = 2^k - 1$, assume for simplicity that the first "k" sets of \mathcal{A}_2 are the "k" one element subsets of $\{1, 2, \dots, k\}$. Thus for $j = 1, 2, \dots, k$ action S_j means: "say population j is the best;" and for $j = k+1, \dots, p$ action S_j means: "say the best is in the subset S_j ." For this problem we assume that the loss function is of the form

$$(4.2.1) \quad L(S_j, \underline{\lambda}) = \sum_{q \in S_j} \alpha_{jq} (\lambda_{[k]} - \lambda_q) ; \quad j = 1, 2, \dots, p ,$$

where $\alpha_{jq} \geq 0$ for all j and q , and $\lambda_{[k]} = \max_{1 \leq j \leq k} \lambda_j$. For the problem of selecting the best of k populations we will use the linear loss structure

$$(4.2.2) \quad L(d_i, \underline{\lambda}) = \lambda_{[k]} - \lambda_i , \quad i = 1, 2, \dots, k .$$

Suppose that r observations are taken from each of the k populations giving us an $r \times k$ observation matrix $\underline{x}^* = (x_1^*, x_2^*, \dots, x_k^*) \in \mathcal{X}^{kr}$ and with the corresponding random matrix $\underline{X}^* = (X_1^*, X_2^*, \dots, X_k^*)$. Then

$$(4.2.3) \quad f_{\underline{\lambda}}(\underline{x}^*) = \prod_{j=1}^k f_{\lambda_j}(x_j^*) \quad \text{where} \quad f_{\lambda_j}(x_j^*) = \prod_{\ell=1}^r f_{\lambda_j}(x_j^{(\ell)}) ,$$

$x_j^{(\ell)}$ being the ℓ 'th observation on X_j ,

and, analogous to (4.1.4), we have

$$(4.2.4) \quad \phi_G(\delta, \underline{x}^*) = \int_{\Omega^k} L(\delta(\underline{x}^*), \underline{\lambda}) f_{\underline{\lambda}}(\underline{x}^*) d G(\underline{\lambda}) .$$

If

$$(4.2.5) \quad c_q = \int_{\Omega^k} (\lambda_{[k]}^{-\lambda_q}) f_{\underline{\lambda}}(\underline{x}^*) d G(\underline{\lambda}) ,$$

then from (4.2.4) and using loss function (4.2.1) we see that

$$\phi_G(S_j, \underline{x}^*) = \sum_{q \in S_j} \alpha_{jq} c_q .$$

We now prove

Theorem 6: In the loss function given by (4.2.1), let $\alpha_{jq} = \alpha > 0$
for $j = 1, 2, \dots, k$. For c_q given by (4.2.5), let $c_{[1]} = \min_{1 \leq q \leq k} c_q$.
If

$$\sum_{q \in S_j} \alpha_{jq} \geq \alpha$$

for every $j = 1, 2, \dots, p$, then $\min_{1 \leq j \leq p} \phi_G(S_j, \underline{x}^*) = \min_{1 \leq j \leq k} \phi_G(S_j, \underline{x}^*)$.

Proof: If

$$\sum_{q \in S_j} \alpha_{jq} \geq \alpha$$

for every $j = 1, 2, \dots, p$, then

$$\sum_{q \in S_j} \alpha_{jq} c_q \geq \sum_{q \in S_j} \alpha_{jq} c_{[1]} \geq \alpha c_{[1]}$$

which implies that

$$(4.2.6) \quad \min_{1 \leq i \leq k} \phi_G(S_i, \underline{x}^*) \leq \phi_G(S_j, \underline{x}^*) \quad \text{for every } j = 1, 2, \dots, p$$

since $\alpha c_{[1]} = \min_{1 \leq i \leq k} \phi_G(S_i, \underline{x}^*)$ and

$$\phi_G(S_j, \underline{x}^*) = \sum_{q \in S_j} \alpha_{jq} c_q.$$

Since (4.2.6) is true for every $j = 1, 2, \dots, p$, $\min_{1 \leq i \leq k} \phi_G(S_i, \underline{x}^*) \leq \min_{1 \leq j \leq p} \phi_G(S_j, \underline{x}^*)$. Since $\min_{1 \leq i \leq k} \phi_G(S_i, \underline{x}^*) \geq \min_{1 \leq j \leq p} \phi_G(S_j, \underline{x}^*)$ also, $\min_{1 \leq j \leq p} \phi_G(S_j, \underline{x}^*) = \min_{1 \leq j \leq k} \phi_G(S_j, \underline{x}^*)$ and the proof is complete.

The following corollary is the main result of this section.

Corollary 3. In the loss function given in (4.2.1), let $\alpha_{jq} = \alpha > 0$ for $j = 1, 2, \dots, k$. If

$$\sum_{q \in S_j} \alpha_{jq} \geq \alpha,$$

then a decision procedure δ_G which is Bayes with respect to G when

$\mathcal{D} = \mathcal{D}_1 = \{d_1, d_2, \dots, d_k\}$ and $L(d_i, \underline{\lambda}) = \alpha(\lambda_{[k]} - \lambda_i)$ for $i = 1, 2, \dots, k$

is also Bayes with respect to G when $\mathcal{D} = \mathcal{D}_2 = \{S_1, S_2, \dots, S_p\}$,

$p = 2^k - 1$, and $L(S_j, \underline{\lambda})$ is given by (4.2.1).

Proof. If δ_G is the Bayes decision procedure with respect to G for \mathcal{D}_1 , then by (4.1.7)

$$(4.2.7) \quad \phi_G(\delta_G, \underline{x}^*) = \min_{1 \leq j \leq k} \{\phi_G(d_j, \underline{x}^*)\}$$

for each \underline{x}^* where $\phi_G(\delta_G, \underline{x}^*)$ is given by (4.2.4). But since δ_G must select some one of the first k sets S_1, S_2, \dots, S_k ,

$$\phi_G(\delta_G, \underline{x}^*) = \min_{1 \leq j \leq k} \phi(S_j, \underline{x}^*) .$$

Thus from Theorem 6

$$\phi_G(\delta_G, \underline{x}^*) = \min_{1 \leq j \leq p} \phi_G(S_j, \underline{x}^*) ,$$

which implies from (4.1.7) that the decision function δ_G is Bayes with respect to G for \mathcal{D}_2 , and the proof is complete.

Therefore the Bayes and empirical Bayes procedures which will be derived for selecting the best of k populations are also Bayes and empirical Bayes procedures for selecting a subset which contains the best population, provided the conditions of Corollary 3 are satisfied. Thus only one population is selected even in the subset formulation.

4.3. Bayes Procedures When G Belongs to a Particular Parametric Family.

Suppose that the a priori distribution G belongs to some particular parametric family \mathcal{G} . \mathcal{G} may be, for example, the class of all normal distributions. If we have r observations on each

population, let $\underline{X}_1^*, \underline{X}_2^*, \dots, \underline{X}_n^*$ be the prior observations of the random matrix \underline{X}^* . For $G \in \mathcal{G}$ and with parameter $\underline{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$, we wish to find an estimate G_n of G , or, equivalently, an estimate $\underline{\theta}_n$ of $\underline{\theta}$ such that the Bayes procedure with respect to G_n , δ_{G_n} , will also be an empirical Bayes procedure with respect to G . Since $G_n \in \mathcal{G}$, finding δ_{G_n} amounts to finding the Bayes procedure with respect to G . We now state

Theorem 7: Let $\underline{X}_1^*, \underline{X}_2^*, \dots$, be independent random matrices, each consisting of $r \times k$ components which are independent random variables. Let G be a k -dimensional distribution function on Ω^k which is independent of the \underline{X}^* 's, and let G_n be a distribution function on Ω^k which is a function of $\underline{X}_1^*, \underline{X}_2^*, \dots, \underline{X}_n^*$ and such that

$$(4.3.1) \quad P\left\{ \lim_{n \rightarrow \infty} G_n(\underline{\lambda}) = G(\underline{\lambda}) \text{ for every continuity point } \underline{\lambda} \text{ of } G \right\} = 1,$$

where probability P is taken with respect to $\underline{X}_1^*, \underline{X}_2^*, \dots$. Let the loss function $L(d_i, \underline{\lambda})$ and densities $f_{\lambda_i}(x_i^*)$ defined in (4.2.3) be such that $L(d_i, \underline{\lambda}) f_{\lambda_i}(x_i^*)$ is bounded and continuous in $\underline{\lambda}$ for every $d_i \in \mathcal{D}_1$, $i = 1, 2, \dots, k$ and $\underline{x}^* \in \mathcal{X}^{kr}$. Let

$$(4.3.2) \quad L(\underline{\lambda}) = \max_{1 \leq i \leq k} \{L(d_i, \underline{\lambda})\}.$$

Then the sequence $\{\delta_{G_n}\}$ is asymptotically optimal with respect to G if

$$(4.3.3) \quad \int_{\Omega^k} L(\underline{\lambda}) dG(\underline{\lambda}) < \infty.$$

Proof. Define

$$(4.3.4) \quad \Delta_G(d_i, \underline{x}^*) = \int_{\Omega^k} [L(d_i, \underline{\lambda}) - L(d_1, \underline{\lambda})] f_{\underline{\lambda}}(\underline{x}^*) dG(\underline{\lambda}),$$

and

$$(4.3.5) \quad \Delta_{G_n}(d_i, \underline{x}^*) = \int_{\Omega^k} [L(d_i, \underline{\lambda}) - L(d_1, \underline{\lambda})] f_{\underline{\lambda}}(\underline{x}^*) dG_n(\underline{\lambda}).$$

If G_n is such that (4.3.1) is true, we have from the Helly-Bray theorem that

$$(4.3.6) \quad \Delta_{G_n}(d_i, \underline{x}^*) \xrightarrow{P} \Delta_G(d_i, \underline{x}^*)$$

since $L(d_i, \underline{\lambda}) f_{\underline{\lambda}}(\underline{x}^*)$ is bounded and continuous in $\underline{\lambda}$ for every $d_i \in \mathcal{D}_1$ and $\underline{x}^* \in \mathcal{X}^{kr}$. However it follows from Corollary 1 (with appropriate notational changes) that if

$$(4.3.7) \quad \Delta_{i,n}(\underline{x}^*) = \Delta_{i,n}(X_1^*, X_2^*, \dots, X_n^*; \underline{x}^*)$$

is a function of the prior observations such that for a.e. \underline{x}^* ,

$$(4.3.8) \quad \Delta_{i,n}(\underline{x}^*) \xrightarrow{P} \Delta_G(d_i, \underline{x}^*)$$

for $i = 1, 2, \dots, k$, then the decision procedure $D = \{\delta_n\}$ defined by

$$(4.3.9) \quad \delta_n(\underline{x}^*) = d_j \text{ where } j \text{ is any positive integer } 1, 2, \dots, k$$

such that $\Delta_{j,n}(\underline{x}^*) = \min_{1 \leq i \leq k} \{\Delta_{i,n}(\underline{x}^*)\}$

is asymptotically optimal relative to G if

$$(4.3.10) \quad \int_{\Omega^k} L(\underline{\lambda}) d G(\underline{\lambda}) < \infty .$$

Replacing (4.3.7) by (4.3.5) it follows from (4.3.6) that the decision procedure $D = \{\delta_{G_n}\}$ defined by

$$(4.3.11) \quad \delta_{G_n}(\underline{x}^*) = d_j \text{ where } j \text{ is any positive integer } 1, 2, \dots, k \text{ such that } \Delta_{G_n}(d_j, \underline{x}^*) = \min_{1 \leq i \leq k} \{\Delta_{G_n}(d_i, \underline{x}^*)\}$$

is asymptotically optimal with respect to G if

$$\int_{\Omega^k} L(\underline{\lambda}) d G(\underline{\lambda}) < \infty$$

and the proof is complete.

Since the j for which

$$\Delta_{G_n}(d_j, \underline{x}^*) = \min_{1 \leq i \leq k} \{\Delta_{G_n}(d_i, \underline{x}^*)\}$$

is the same as the j for which

$$\phi_{G_n}(d_j, \underline{x}^*) = \min_{1 \leq i \leq k} \{\phi_{G_n}(d_i, \underline{x}^*)\}$$

where

$$\phi_{G_n}(d_i, \underline{x}^*) = \int_{\Omega^k} L(d_i, \underline{\lambda}) f_{\underline{\lambda}}(\underline{x}^*) d G_n(\underline{\lambda}) ,$$

it follows from (4.1.7) that δ_{G_n} is a Bayes procedure with respect to G_n . δ_{G_n} is also an empirical Bayes procedure with respect to G due to (4.1.14).

If $G \in \mathcal{G}$, some parametric family with parameter $\underline{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$, then Theorem 7 suggests the following procedure:

- (1) For specified \mathcal{G} find the Bayes procedure δ_G as a function of the observation \underline{x}^* and the parameter $\underline{\theta}$ of G ; i.e.

$$\delta_G(\underline{x}^*) = h(\underline{x}^*, \underline{\theta}).$$

- (2) Verify that $L(d_i, \underline{\lambda}) f_{\underline{\lambda}}(\underline{x}^*)$ is bounded and continuous in $\underline{\lambda}$ for any $d_i \in \mathcal{D}_1$. Verify that

$$\int_{\Omega^k} L(\underline{\lambda}) d G(\underline{\lambda}) < \infty.$$

- (3) Find an estimate $\underline{\theta}_n$ of $\underline{\theta}$ such that $G_n \rightarrow G$ with probability one. Then $\delta_{G_n}(\underline{x}^*) = h(\underline{x}^*, \underline{\theta}_n)$ is an empirical Bayes procedure with respect to $G \in \mathcal{G}$.

From (4.1.7) it follows, again, that the Bayes decision procedure δ_G with respect to the a priori distribution

$$G = \prod_{j=1}^k G_j$$

minimizes

$$(4.3.12) \quad \phi_G(d_i, \underline{x}^*) = \int_{\Omega^k} L(d_i, \underline{\lambda}) f_{\underline{\lambda}}(\underline{x}^*) d G(\underline{\lambda}) \quad \text{for } d_i \in \mathcal{D}_1,$$

where $i = 1, 2, \dots, k$. Using the loss function (4.2.2) we have

$$(4.3.13) \quad \phi_G(d_i, \underline{x}^*) = \int_{\Omega^k} \lambda_{[k]} f_{\underline{\lambda}}(\underline{x}^*) dG(\underline{\lambda}) - \int_{\Omega^k} \lambda_i f_{\underline{\lambda}}(\underline{x}^*) dG(\underline{\lambda}) .$$

Since the first term in (4.3.13) is independent of i , the Bayes procedure is defined by

$$(4.3.14) \quad \delta_G(\underline{x}^*) = d_j \text{ where } j \text{ is any positive integer } 1, 2, \dots, k$$

$$\text{such that } \int_{\Omega^k} \lambda_j f_{\underline{\lambda}}(\underline{x}^*) dG(\underline{\lambda}) = \max_{1 \leq i \leq k} \left\{ \int_{\Omega^k} \lambda_i f_{\underline{\lambda}}(\underline{x}^*) dG(\underline{\lambda}) \right\} .$$

Suppose that G_i has a density function $g_i(\lambda_i)$ with respect to some measure ν_i . Then if

$$(4.3.15) \quad f_{G_j}(x_j^*) = \int_{\Omega} f_{\lambda_j}(x_j^*) dG_j(\lambda_j)$$

and

$$(4.3.16) \quad g_i(\lambda_i | x_i^*) = \frac{f_{\lambda_i}(x_i^*) g_i(\lambda_i)}{f_{G_i}(x_i^*)} ,$$

$$\begin{aligned} (4.3.17) \quad \int_{\Omega^k} \lambda_i f_{\underline{\lambda}}(\underline{x}^*) dG(\underline{\lambda}) &= \int_{\Omega^k} \lambda_i \left(\prod_{j=1}^k f_{\lambda_j}(x_j^*) \right) d \left(\prod_{j=1}^k G_j \right) \\ &= \left\{ \int_{\Omega} \lambda_i f_{\lambda_i}(x_i^*) g_i(\lambda_i) d\nu_i(\lambda_i) \right\} \left\{ \prod_{\substack{j=1 \\ j \neq i}}^k f_{G_j}(x_j^*) \right\} \\ &= \left\{ \left[\int_{\Omega} \lambda_i g_i(\lambda_i | x_i^*) d\nu_i(\lambda_i) \right] [f_{G_i}(x_i^*)] \right\} \left\{ \prod_{\substack{j=1 \\ j \neq i}}^k f_{G_j}(x_j^*) \right\} \\ &= E(\Lambda_i | x_i^*) \cdot f_G(\underline{x}^*) , \end{aligned}$$

where

$$(4.3.18) \quad E(\Lambda_i | x_i^*) = \int_{\Omega} \lambda_i g_i(\lambda_i | x_i^*) d\nu_i(\lambda_i)$$

is the a posteriori mean for the i 'th population, and

$$(4.3.19) \quad f_G(\underline{x}^*) = \prod_{j=1}^k f_{G_j}(x_j^*) .$$

Since $f_G(\underline{x}^*)$ is independent of i , we have from (4.3.14) that the largest a posteriori mean gives the Bayes procedure with respect to G .

We have proved

Theorem 8: From each of k populations let there be r observations taken on a random variable with density $f_{\lambda_i}(x)$ for $i = 1, 2, \dots, k$. If λ_i is distributed according to the density $g_i(\lambda_i)$ with respect to some measure ν_i , then the Bayes procedure for selecting the best population under the linear loss function (4.2.2) is given by

$$(4.3.20) \quad \delta_G(\underline{x}^*) = d_j \text{ where } j \text{ is any integer } 1, 2, \dots, k \text{ such that } E(\Lambda_j | x_j^*) = \max_{1 \leq i \leq k} \{E(\Lambda_i | x_i^*)\} .$$

Calculations may often be simplified if we use sufficient statistics. Using the usual factorization criteria we see that $t_j = t_j(x_j^*)$ is sufficient for $f_{\lambda_j}(x_j^*)$ if

$$(4.3.21) \quad f_{\lambda_j}(x_j^*) = h_{\lambda_j}(t_j) p(x_j^*) ,$$

where h depends on the observations only through t_j , and p does not involve the parameter λ_j . We then have the following sufficiency lemma.

Lemma 2. Suppose $f_{\lambda_j}(x_j^*)$ admits a sufficient statistic $t_j = t_j(x_j^*)$ for $j = 1, 2, \dots, k$. Then

$$\int_{\Omega^k} \lambda_j f_{\lambda_j}(x_j^*) dG(\underline{\lambda}) = \max_{1 \leq i \leq k} \left\{ \int_{\Omega^k} \lambda_i f_{\lambda_i}(x_i^*) dG(\underline{\lambda}) \right\}$$

if and only if

$$(4.3.22) \quad \int_{\Omega^k} \lambda_j h_{\lambda_j}(\underline{t}) dG(\underline{\lambda}) = \max_{1 \leq i \leq k} \left\{ \int_{\Omega^k} \lambda_i h_{\lambda_i}(\underline{t}) dG(\underline{\lambda}) \right\},$$

where $h_{\lambda_j}(\underline{t}) = \prod_{j=1}^k h_{\lambda_j}(t_j)$.

Proof. Since $f_{\lambda_j}(x_j^*) = h_{\lambda_j}(t_j) p(x_j^*)$ from (4.3.21), we have

$$\begin{aligned} f_{\lambda_j}(x_j^*) &= \prod_{j=1}^k f_{\lambda_j}(x_j^*) = \prod_{j=1}^k h_{\lambda_j}(t_j) p(x_j^*) \\ (4.3.23) \quad &= \left\{ \prod_{j=1}^k h_{\lambda_j}(t_j) \right\} \left\{ \prod_{j=1}^k p(x_j^*) \right\} \\ &= h_{\lambda_j}(\underline{t}) p(\underline{x}^*), \end{aligned}$$

where $p(\underline{x}^*) = \prod_{j=1}^k p(x_j^*)$ is independent of $\underline{\lambda}$ and i . Thus

$$\begin{aligned} & \max_{1 \leq i \leq k} \int_{\Omega^k} \lambda_i f_{\underline{\lambda}}(\underline{x}^*) d G(\underline{\lambda}) \\ &= p(\underline{x}^*) \max_{1 \leq i \leq k} \left\{ \int_{\Omega^k} \lambda_i h_{\underline{\lambda}}(\underline{t}) d G(\underline{\lambda}) \right\} \end{aligned}$$

from which the lemma follows.

Thus if we replace $E(\Lambda_j | \underline{x}_j^*)$ by $E(\Lambda_j | t_j)$ in Theorem 8, the computations are often reduced.

4.4. Empirical Bayes Procedures when G Belongs to a Particular Parametric Family.

In order to find the empirical Bayes procedures for selecting the best of k populations, we require the following lemmas.

Lemma 3: Let G_j be a one-dimensional distribution function, and suppose that $G_{n,j}$, $n = 1, 2, \dots$, is a sequence of distribution functions converging to G_j at the points of continuity of G_j for $j = 1, 2, \dots, k$. Then

$$G_{\pi,n} = \prod_{j=1}^k G_{n,j} \text{ converges to } G = \prod_{j=1}^k G_j \text{ at the points of continuity}$$

of G .

Proof: Let $\underline{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_k)$ be a point of continuity of G . If $G(\underline{\lambda}) \neq 0$, then λ_j is a point of continuity of G_j for $j = 1, 2, \dots, k$, and by hypothesis $\prod_{j=1}^k G_{n,j} \rightarrow \prod_{j=1}^k G_j$ at $\underline{\lambda}$. If $G(\underline{\lambda}) = 0$, then

$G_j(\lambda_j) = 0$ for some j ; i.e. for $j \in K$, a subset of $\{1, 2, \dots, k\}$

which is non-empty. Also, for some j in K , λ_j must be a continuity point of G_j for otherwise $\underline{\lambda}$ would not be a continuity point of G . Thus there exists a sequence $\{G_{n,j}\}$ such that $G_{n,j}(\lambda_j) \rightarrow G_j(\lambda_j) = 0$. Therefore $G_{\pi,n} \rightarrow G$ at $\underline{\lambda}$, and the proof is complete.

Lemma 4: Let X_1^* , X_2^* , ..., be independent $k \times r$ -dimensional random matrices consisting of kr components which are independent random variables. Suppose that $G_{n,j}$ is a distribution function on Ω which is a function of $X_{1,j}^*$, $X_{2,j}^*$, ..., $X_{n,j}^*$ such that $G_{n,j}$ converges to G_j , a distribution function on Ω , with probability one for $j = 1, 2, \dots, k$ with probability being taken with respect to $X_{1,j}^*$, $X_{2,j}^*$, Then

$G_{\pi,n} = \prod_{j=1}^k G_{n,j}$ converges to $G = \prod_{j=1}^k G_j$ with probability one where probability here is the product of the above probabilities.

Proof: Let $B_j = \{y = (x_{1,j}^*, x_{2,j}^*, \dots) : G_{n,j} \rightarrow G_j \text{ at } y\}$. Then

by hypothesis, $P\{B_j\} = 1$ for $j = 1, 2, \dots, k$. Using Lemma 3,

$$G_{\pi,n} = \prod_{j=1}^k G_{n,j} \rightarrow G = \prod_{j=1}^k G_j \text{ for } y \in B = \prod_{j=1}^k B_j. \text{ Thus } P\{B\} = \prod_{j=1}^k P\{B_j\} = 1$$

which completes the proof.

Thus if an estimate $\theta_{n,j}$ of the parameter θ_j of G_j is available such that the distribution function corresponding to $\theta_{n,j}$, $G_{n,j}$, converges to G_j with probability one, then by Lemma 4

$G_{\pi,n} = \prod_{j=1}^k G_{n,j}$ converges to $G = \prod_{j=1}^k G_j$ with probability one. There-

fore by Theorem 7 an empirical Bayes procedure with respect to G is

obtained provided that $L(d_i, \underline{\lambda}) f_{\underline{\lambda}}(\underline{x}^*)$ is bounded and continuous in $\underline{\lambda}$ for each \underline{x}^* and any $d_i \in \mathcal{D}_1$, $i = 1, 2, \dots, k$ and that

$$\int_{\Omega^k} L(\underline{\lambda}) dG(\underline{\lambda}) < \infty.$$

If $\lambda_{[1]} = \min_{1 \leq i \leq k} \lambda_i$, then since $L(\underline{\lambda}) = \max \{L(d_i, \underline{\lambda}), d_i \in \mathcal{D}_1\}$,

we have, using loss function (4.2.2),

$$\begin{aligned} \int_{\Omega^k} L(\underline{\lambda}) dG(\underline{\lambda}) &= \int_{\Omega^k} \max_{1 \leq i \leq k} \{\lambda_{[k]} - \lambda_i\} dG(\underline{\lambda}) \\ &= \int_{\Omega^k} (\lambda_{[k]} - \lambda_{[1]}) dG(\underline{\lambda}) \\ &\leq \int_{\Omega^k} \left\{ \sum_{i=1}^k |\lambda_i| \right\} dG(\underline{\lambda}) \\ (4.4.1) \quad &= \int_{\Omega} \dots \int_{\Omega} \left\{ \sum_{i=1}^k |\lambda_i| \right\} dG_1 \dots dG_k \\ &= \sum_{i=1}^k \left(\int_{\Omega} |\lambda_i| dG_i \right) \left(\int_{\Omega} \dots \int_{\Omega} \prod_{\substack{j=1 \\ j \neq i}}^k dG_j \right) \\ &= \sum_{i=1}^k \int_{\Omega} |\lambda_i| dG_i \\ &= \sum_{i=1}^k E(|\lambda_i|). \end{aligned}$$

Therefore

$$(4.4.2) \quad \int_{\Omega^k} L(\underline{\lambda}) dG(\underline{\lambda}) < \infty \quad \text{if } E(|\lambda_i|) < \infty \quad \text{for all } i = 1, 2, \dots, k.$$

The Bayes procedure for selecting the best of k populations under linear loss structure (4.2.2) is given by (4.3.20). If G is unknown but exists, and prior observations $\underline{X}_1^*, \underline{X}_2^*, \dots, \underline{X}_n^*$ are available, then using Lemmas 3 and 4 and Theorem 7 it follows that an empirical Bayes procedure for selecting the best of k populations is given by

$$(4.4.3) \quad \delta_{G_{\pi,n}}(\underline{x}^*) = d_j \quad \text{where } j \text{ is any integer } 1, 2, \dots, k$$

$$\text{such that } E_n(\Lambda_j | \underline{x}_j^*) = \max_{1 \leq i \leq k} \{E_n(\Lambda_i | \underline{x}_i^*)\}$$

where E_n denotes expectation with respect to $G_{\pi,n}$, provided an estimate $E_n(\Lambda_j | \underline{x}_j^*)$ of $E(\Lambda_j | \underline{x}_j^*)$ is available such that $G_{n,j}$, the distribution function corresponding to $E_n(\Lambda_j | \underline{x}_j^*)$, converges to G_j at the points of continuity of G_j with probability one, and the conditions of Theorem 7 are met.

4.5. Examples

As an example of the above technique consider the "normal-normal" case where $f_{\lambda_i}(x_i)$ has a normal density with unknown mean λ_i and known variance σ_i^2 , and G_i has a normal density $g_i(\lambda_i)$ with unknown mean θ_i and known variance β_i^2 . We wish to compute

$$E(\Lambda_i | \underline{x}_i^*) = \int \lambda_i g_i(\lambda_i | \underline{x}_i^*) d\nu_i(\lambda_i),$$

where

$$g_i(\lambda_i | x_i^*) = \frac{f_{\lambda_i}(x_i^*) g_i(\lambda_i)}{f_{G_i}(x_i^*)} .$$

Since $f_{\lambda_i}(x_i)$ is $N(\lambda_i, \sigma_i^2)$ and the r observations per population are independent, we have for $i = 1, 2, \dots, k$

$$\begin{aligned} f_{\lambda_i}(x_i^*) &= \prod_{\ell=1}^r f_{\lambda_i}(x_i^{(\ell)}) \\ &= (2\pi\sigma_i^2)^{-r/2} \exp \left\{ -\frac{1}{2\sigma_i^2} \sum_{\ell=1}^r (x_i^{(\ell)} - \lambda_i)^2 \right\} \\ (4.5.1) \quad &= (2\pi\sigma_i^2)^{-r/2} \exp \left\{ -\frac{1}{2\sigma_i^2} \left[r \left(\frac{1}{r} \sum_{\ell=1}^r (x_i^{(\ell)} - \bar{x}_i)^2 \right) + r(\bar{x}_i - \lambda_i)^2 \right] \right\} \\ &= (2\pi\sigma_i^2)^{-r/2} \exp \left\{ -\frac{1}{2\sigma_i^2} [r s_i^2 + r(\bar{x}_i - \lambda_i)^2] \right\} , \end{aligned}$$

where

$$(4.5.2) \quad s_i^2 = \frac{1}{r} \sum_{\ell=1}^r (x_i^{(\ell)} - \bar{x}_i)^2 ,$$

and

$$(4.5.3) \quad \bar{x}_i = \frac{1}{r} \sum_{\ell=1}^r x_i^{(\ell)} .$$

Since $g_i(\lambda_i)$ is $N(\theta_i, \beta_i^2)$ with respect to some measure ν_i , we have using (4.5.1)

$$\begin{aligned}
 f_{G_i}(x_i^*) &= \int_{-\infty}^{\infty} \left\{ (2\pi\sigma_i^2)^{-(r/2)} \exp \left(-\frac{1}{2\sigma_i^2} [rs_i^2 + r(\bar{x}_i - \lambda_i)^2] \right) \right\} \\
 &\quad \cdot \left\{ (2\pi\beta_i^2)^{-(1/2)} \exp \left(-\frac{1}{2\beta_i^2} (\lambda_i - \theta_i)^2 \right) \right\} d\nu_i(\lambda_i) \\
 (4.5.4) \quad &= \left\{ (2\pi)^{-\frac{r+1}{2}} \sigma_i^{-r} \beta_i^{-1} \right\} \left\{ \int_{-\infty}^{\infty} \exp \left[-\frac{rs_i^2 + r(\bar{x}_i - \lambda_i)^2}{2\sigma_i^2} - \frac{(\lambda_i - \theta_i)^2}{2\beta_i^2} \right] d\nu_i(\lambda_i) \right\} \\
 &= \left\{ (2\pi)^{-\frac{r+1}{2}} \sigma_i^{-r} \beta_i^{-1} \exp \left(-\frac{rs_i^2}{2\sigma_i^2} \right) \right\} \left\{ \int_{-\infty}^{\infty} \exp \left[-\frac{1}{2} \frac{(\bar{x}_i - \lambda_i)^2}{\frac{\sigma_i^2}{r}} \right. \right. \\
 &\quad \left. \left. \cdot \exp \left(-\frac{1}{2} \left(\frac{\lambda_i - \theta_i}{\beta_i} \right)^2 \right) d\nu_i(\lambda_i) \right] \right\} .
 \end{aligned}$$

Now

$$\begin{aligned}
 (4.5.5) \quad &\left\{ \int_{-\infty}^{\infty} \exp \left[-\frac{1}{2} \frac{(\bar{x}_i - \lambda_i)^2}{\frac{\sigma_i^2}{r}} \right] \exp \left[-\frac{1}{2} \left(\frac{\lambda_i - \theta_i}{\beta_i} \right)^2 \right] d\nu_i(\lambda_i) \right\} \\
 &= \int_{-\infty}^{\infty} \exp \left[-\frac{1}{2} \left[\frac{(\bar{x}_i - \lambda_i)^2}{\frac{\sigma_i^2}{r}} + \left(\frac{\lambda_i - \theta_i}{\beta_i} \right)^2 \right] \right] d\nu_i(\lambda_i) ,
 \end{aligned}$$

and

$$\begin{aligned}
 & \frac{(\bar{x}_i - \lambda_i)^2}{\sigma_i^2/r} + \frac{(\lambda_i - \theta_i)^2}{\beta_i^2} = \lambda_i^2 \left[\frac{1}{\sigma_i^2/r} + \frac{1}{\beta_i^2} \right] \\
 & - 2\lambda_i \left[\frac{\bar{x}_i}{\sigma_i^2/r} + \frac{\theta_i}{\beta_i^2} \right] + \frac{\bar{x}_i^2}{\sigma_i^2/r} + \frac{\theta_i^2}{\beta_i^2} \\
 (4.5.6) \quad & = \left(\lambda_i \sqrt{\frac{1}{\sigma_i^2/r} + \frac{1}{\beta_i^2}} - \frac{\frac{\bar{x}_i}{\sigma_i^2/r} + \frac{\theta_i}{\beta_i^2}}{\sqrt{\frac{1}{\sigma_i^2/r} + \frac{1}{\beta_i^2}}} \right)^2 \\
 & - \frac{\left[\frac{\bar{x}_i}{\sigma_i^2/r} + \frac{\theta_i}{\beta_i^2} \right]^2}{\frac{1}{\sigma_i^2/r} + \frac{1}{\beta_i^2}} + \frac{\bar{x}_i^2}{\sigma_i^2/r} + \frac{\theta_i^2}{\beta_i^2} .
 \end{aligned}$$

Let

$$A = \sqrt{\frac{1}{\sigma_i^2/r} + \frac{1}{\beta_i^2}} , \quad B = \frac{\frac{\bar{x}_i}{\sigma_i^2/r} + \frac{\theta_i}{\beta_i^2}}{\sqrt{\frac{1}{\sigma_i^2/r} + \frac{1}{\beta_i^2}}} ,$$

and

$$C = \frac{\bar{x}_i^2}{\sigma_i^2/r} + \frac{\theta_i^2}{\beta_i^2} - \frac{\left[\frac{\bar{x}_i}{\sigma_i^2/r} + \frac{\theta_i}{\beta_i^2} \right]^2}{\frac{1}{\sigma_i^2/r} + \frac{1}{\beta_i^2}} = \frac{(\bar{x}_i - \theta_i)^2}{\beta_i^2 + \sigma_i^2/r} .$$

Substituting (4.5.6) in (4.5.5) we get

$$\begin{aligned}
 & \int_{-\infty}^{\infty} \exp - \frac{1}{2} [(A\lambda_i - B)^2 + C] d\nu_i(\lambda_i) \\
 (4.5.7) \quad & = e^{-\frac{C}{2}} \frac{\sqrt{2\pi}}{A} \\
 & = \sqrt{2\pi} \frac{\sigma_i \beta_i}{\sqrt{r \beta_i^2 + \sigma_i^2}} \exp - \frac{1}{2} \left(\frac{(\bar{x}_i - \theta_i)^2}{\beta_i^2 + \frac{\sigma_i^2}{r}} \right)
 \end{aligned}$$

since

$$\int_{-\infty}^{\infty} \exp - \frac{1}{2} (A\lambda_i - B)^2 d\nu_i(\lambda_i) = \frac{\sqrt{2\pi}}{A} .$$

Substituting (4.5.7) in (4.5.4) we then get

$$\begin{aligned}
 f_{G_i}(x_i^*) &= \left\{ (2\pi)^{-\frac{r+1}{2}} \sigma_i^{-r} \beta_i^{-1} \exp \left(\frac{-rs_i^2}{2\sigma_i^2} \right) \right\} \\
 (4.5.8) \quad & \cdot \left\{ \sqrt{2\pi} \frac{\sigma_i \beta_i}{(\sigma_i^2 + r\beta_i^2)^{1/2}} \exp \left(-\frac{1}{2} \frac{(\bar{x}_i - \theta_i)^2}{\frac{\sigma_i^2}{r} + \beta_i^2} \right) \right\} \\
 &= \left\{ \frac{(2\pi)^{-\frac{r}{2}} \sigma_i^{-r+1}}{(\sigma_i^2 + r\beta_i^2)^{1/2}} \right\} \left\{ \exp \left[-\frac{rs_i^2}{2\sigma_i^2} - \frac{r(\bar{x}_i - \theta_i)^2}{2(\sigma_i^2 + r\beta_i^2)} \right] \right\} .
 \end{aligned}$$

Thus

$$g_i(\lambda_i | x_i^*) = \frac{f_{\lambda_i}(x_i^*) g_i(\lambda_i)}{f_{G_i}(x_i^*)}$$

$$\begin{aligned}
 & \left\{ (2\pi)^{-\frac{r+1}{2}} \sigma_i^{-r} \beta_i^{-1} \right\} \left\{ \exp \left[-\frac{rs_i^2}{2\sigma_i^2} - \frac{r(\bar{x}_i - \lambda_i)^2}{2\sigma_i^2} - \frac{(\lambda_i - \theta_i)^2}{2\beta_i^2} \right] \right\} \\
 &= \frac{\left\{ (2\pi)^{-\frac{r}{2}} \sigma_i^{-r+1} \right\} \left\{ \exp \left[-\frac{rs_i^2}{2\sigma_i^2} - \frac{r(\bar{x}_i - \theta_i)^2}{2(\sigma_i^2 + r\beta_i^2)} \right] \right\}}{(\sigma_i^2 + r\beta_i^2)^{1/2}} \\
 &= \left\{ (2\pi)^{-\frac{1}{2}} (\sigma_i^2 + r\beta_i^2)^{\frac{1}{2}} (\sigma_i \beta_i)^{-1} \right\} \left\{ \exp \left[-\frac{1}{2} \left(\frac{\sigma_i^2 + r\beta_i^2}{\sigma_i^2 \beta_i^2} \right) \left(\lambda_i - \frac{r\beta_i^2 \bar{x}_i + \sigma_i^2 \theta_i}{\sigma_i^2 + r\beta_i^2} \right)^2 \right] \right\}
 \end{aligned}$$

which is the normal density with mean

$$(4.5.9) \quad \int_{-\infty}^{\infty} \lambda_i g_i(\lambda_i | x_i^*) d\nu_i(\lambda_i) = \frac{r\beta_i^2 \bar{x}_i + \sigma_i^2 \theta_i}{\sigma_i^2 + r\beta_i^2} = E(\Lambda_i | x_i^*) .$$

If we assume that θ_i is finite, then by (4.4.2),

$$\int_{\Omega^k} L(\underline{\lambda}) dG(\underline{\lambda}) < \infty .$$

Also,

$$\begin{aligned}
 (4.5.10) \quad L(d_i, \underline{\lambda}) f_{\underline{\lambda}}(\underline{x}^*) &= (\lambda_{[k]} - \lambda_i) \prod_{j=1}^k f_{\lambda_j}(x_j^*) \\
 &= M(\lambda_{[k]} - \lambda_i) \exp \left\{ -\frac{1}{2} \sum_{j=1}^k \left(\frac{\bar{x}_j - \lambda_j}{\sigma_j} \right)^2 \right\} ,
 \end{aligned}$$

where

$$M = \prod_{j=1}^k (2\pi\sigma_j^2)^{-\frac{r}{2}} \exp \left(-\frac{rs_j^2}{2\sigma_j^2} \right)$$

is constant with respect to $\underline{\lambda}$ for fixed \underline{x}_j^* and where

$$s_j^2 = \frac{1}{r} \sum_{\ell=1}^r (x_j^{(\ell)} - \bar{x}_j)^2$$

from before. Then (4.5.10) is bounded and continuous in $\underline{\lambda}$ for given \underline{x}^* and any $d_i \in \mathcal{D}_1$.

If $r = 1$ in (4.5.8), it follows that the random variable X_j has a normal unconditional density with mean θ_j and variance $\sigma_j^2 + \beta_j^2$ and thus the nr prior independent observations on X_j provide a suitable estimate of θ_j . Define

$$(4.5.11) \quad \theta_{n,j} = \bar{x}_j = \frac{1}{nr} \left(\sum_{i=1}^n \sum_{\ell=1}^r x_{ij}^{(\ell)} \right),$$

the overall average of nr observations on X_j where $j = 1, 2, \dots, k$. By the strong law of large numbers $\theta_{n,j} \xrightarrow{\text{a.s.}} \theta_j$ so that defining $G_{n,j}$ by G_j with θ_j replaced by $\theta_{n,j}$, we have that $G_{n,j} \rightarrow G_j$ with probability one. Thus using Theorems 7 and 8, Lemmas 3 and 4, (4.4.3), and (4.5.9) we see that an empirical Bayes procedure for selecting the best of k populations under the linear loss function (4.2.2) is given by

$$(4.5.12) \quad \delta_{G_{\pi,n}}(\underline{x}^*) = d_j \text{ where } j \text{ is any integer } 1, 2, \dots, k \text{ such}$$

$$\text{that } \frac{r\beta_j^2 \bar{x}_j + \sigma_j^2 \bar{x}_j}{\sigma_j^2 + r\beta_j^2} = \max_{1 \leq i \leq k} \left\{ \frac{r\beta_i^2 \bar{x}_i + \sigma_i^2 \bar{x}_i}{\sigma_i^2 + r\beta_i^2} \right\}.$$

In a similar manner Deely [2] derives Bayes and empirical Bayes procedures for selecting the best of k populations for the following cases.

(1) Normal-uniform. $f_{\lambda_j}(x_j)$ is $N(\lambda_j, \sigma_j^2)$ where σ_j^2 is known, and G_j is the distribution function for a uniform density on $(\theta_j - a_j, \theta_j + a_j)$. The Bayes procedure with respect to

$$G = \prod_{j=1}^k G_j$$

where the θ_j 's and a_j 's are known, and

$$\alpha_j = \left(\frac{r}{\sigma_j^2} \right)^{1/2} (\theta_j + a_j - \bar{x}_j), \quad \text{and} \quad \beta_j = \left(\frac{r}{\sigma_j^2} \right)^{1/2} (\theta_j - a_j - \bar{x}_j)$$

where

$$\bar{x}_j = \frac{1}{r} \sum_{\ell=1}^r x_j^{(\ell)}$$

is given by

$$(4.5.13) \quad \delta_G(\underline{x}^*) = d_j \quad \text{where } j \text{ is any integer } 1, 2, \dots, k \text{ such}$$

$$\text{that } \frac{\varphi(\beta_j) - \varphi(\alpha_j)}{\Phi(\alpha_j) - \Phi(\beta_j)} + \frac{r\bar{x}_j}{\sigma_j^2} = \max_{1 \leq i \leq k} \left\{ \frac{\varphi(\beta_i) - \varphi(\alpha_i)}{\Phi(\alpha_i) - \Phi(\beta_i)} + \frac{r\bar{x}_i}{\sigma_i^2} \right\},$$

where $\varphi(u)$ is the standard normal density function, and $\Phi(u)$ is the standard normal distribution function. Assuming that a_j is known, and θ_j is unknown but finite for $j = 1, 2, \dots, k$, the empirical Bayes procedure

based on nr prior observations and which is asymptotically optimal to δ_G is given by

$$(4.5.14) \quad \delta_{G_{\pi,n}}(\underline{x}^*) = d_j \text{ where } j \text{ is any integer } 1, 2, \dots, k \text{ such}$$

$$\text{that } \frac{\varphi(\beta_j^!) - \varphi(\alpha_j^!)}{\Phi(\alpha_j^!) - \Phi(\beta_j^!)} + \frac{r\bar{x}_j}{\sigma_j^2} = \max_{1 \leq i \leq k} \left\{ \frac{\varphi(\beta_i^!) - \varphi(\alpha_i^!)}{\Phi(\alpha_i^!) - \Phi(\beta_i^!)} + \frac{r\bar{x}_i}{\sigma_i^2} \right\},$$

where

$$\alpha_j^! = \left(\frac{r}{\sigma_j^2} \right)^{1/2} (\bar{x}_j + a_j - \bar{x}_j),$$

and

$$\beta_j^! = \left(\frac{r}{\sigma_j^2} \right)^{1/2} (\bar{x}_j - a_j - \bar{x}_j)$$

for $j = 1, 2, \dots, k$, where

$$\bar{x}_j = \frac{1}{nr} \sum_{i=1}^n \sum_{\ell=1}^r x_{ij}^{(\ell)}.$$

(2) Binomial-beta.

$$f_{\lambda_j}(x_j) = \binom{u_j}{x_j} \lambda_j^{x_j} (1 - \lambda_j)^{u_j - x_j},$$

where u_j is the number of trials, x_j is the number of successes, and λ_j is the probability of success on a single trial. G_j is the distribution

function for a beta density $g_j(\lambda_j) = c_j \lambda_j^{\theta_j-1} (1-\lambda_j)^{\nu_j-1}$, where ν_j and θ_j are non-negative parameters, and

$$c_j = \frac{\Gamma(\nu_j + \theta_j)}{\Gamma(\theta_j) \Gamma(\nu_j)}.$$

The Bayes procedure for selecting the best population under the linear loss function (4.2.2) when ν_j and θ_j are known is given by

$$(4.5.15) \quad \delta_G(\underline{x}^*) = d_j \text{ where } j \text{ is any integer } 1, 2, \dots, k \text{ such}$$

$$\text{that } \frac{\bar{x}_j + \frac{1}{r} \theta_j}{u_j + \frac{1}{r} \nu_j} = \max_{1 \leq i \leq k} \left\{ \frac{\bar{x}_i + \frac{1}{r} \theta_i}{u_i + \frac{1}{r} \nu_i} \right\}.$$

Assuming that ν_j is known and θ_j is unknown but finite for $j = 1, 2, \dots, k$, and empirical Bayes procedure based on nr prior observations is given by

$$(4.5.16) \quad \delta_{G, \pi, n}(\underline{x}^*) = d_j \text{ where } j \text{ is any integer } 1, 2, \dots, k \text{ such}$$

$$\text{that } \frac{\bar{x}_j + \frac{1}{r} \frac{\nu_j}{u_j} \bar{\bar{x}}_j}{u_j + \frac{1}{r} \nu_j} = \max_{1 \leq i \leq k} \left\{ \frac{\bar{x}_i + \frac{1}{r} \frac{\nu_i}{u_i} \bar{\bar{x}}_i}{u_i + \frac{1}{r} \nu_i} \right\},$$

where

$$\bar{\bar{x}}_j = \frac{1}{nr} \sum_{i=1}^n \sum_{\ell=1}^r x_{ij}^{(\ell)}.$$

$\delta_{G, \pi, n}$ is asymptotically optimal to δ_G .

(3) Poisson-gamma.

$$f_{\lambda_j}(x_j) = \frac{e^{-\lambda_j} \lambda_j^{x_j}}{x_j!}$$

for $x_j = 0, 1, 2, \dots$, and $\lambda_j > 0$, and $G_j(\lambda_j)$ is the distribution function for the gamma density

$$g_j(\lambda_j) = \frac{\alpha_j^{\theta_j} \lambda_j^{\theta_j-1} e^{-\lambda_j \alpha_j}}{\Gamma(\theta_j)}$$

in which α_j and θ_j are non-negative parameters. Then the Bayes procedure with respect to G for selecting the best population under the linear loss function (4.2.2) and when α_j, θ_j are known, is given by

$$(4.5.17) \quad \delta_G(\underline{x}^*) = d_j \text{ where } j \text{ is any integer } 1, 2, \dots, k \text{ such}$$

$$\text{that } \frac{r\bar{x}_j + \theta_j}{r + \alpha_j} = \max_{1 \leq i \leq k} \left\{ \frac{r\bar{x}_i + \theta_i}{r + \alpha_i} \right\}.$$

Assuming that α_j is known and θ_j is unknown but finite for $j = 1, 2, \dots, k$, the empirical Bayes procedure which is based on nr prior observations and is asymptotically optimal to δ_G is given by

$$(4.5.18) \quad \delta_{G_{\pi, n}}(\underline{x}^*) = d_j \text{ where } j \text{ is any integer } 1, 2, \dots, k \text{ such}$$

$$\text{that } \frac{r\bar{x}_j + \alpha_j \bar{\bar{x}}_j}{r + \alpha_j} = \max_{1 \leq i \leq k} \left\{ \frac{r\bar{x}_i + \alpha_i \bar{\bar{x}}_i}{r + \alpha_i} \right\}.$$

4.6. Empirical Bayes Procedures. The General Case.

Suppose we drop the assumption that the a priori distribution G is a member of some particular parametric family, and assume instead that $G \in \mathcal{G}$ where

$$(4.6.1) \quad \mathcal{G} = \{G : \int_{\Omega^k} L(d_i, \underline{\lambda}) d G(\underline{\lambda}) < \infty \text{ for } d_i \in \mathcal{D}_1\} .$$

If we define $\Delta_G(d_i, \underline{x}^*)$ by (4.3.4) and $\Delta_{i,n}(\underline{x}^*)$ by (4.3.7) such that $\Delta_{i,n}(\underline{x}^*) \xrightarrow{P} \Delta_G(d_i, \underline{x}^*)$, then the sequence of procedures $D = \{\delta_n\}$ defined by (4.3.9) is asymptotically optimal relative to G if

$$\int_{\Omega^k} L(\underline{\lambda}) d G(\underline{\lambda}) < \infty .$$

From (4.4.2) the above condition will be satisfied for $G \in \bar{\mathcal{G}}$ where

$$(4.6.2) \quad \bar{\mathcal{G}} = \{G : G = \prod_{j=1}^k G_j \text{ and } G_j \text{ is a distribution function}$$

$$\text{on } \Omega \text{ such that } \int_{\Omega} \lambda_j d G(\lambda_j) < \infty \text{ for } j = 1, 2, \dots, k\} .$$

Using loss function (4.2.2) we have

$$(4.6.3) \quad \begin{aligned} \Delta_G(d_i, \underline{x}^*) &= \int_{\Omega^k} [L(d_i, \underline{\lambda}) - L(d_1, \underline{\lambda})] f_{\underline{\lambda}}(\underline{x}^*) d G(\underline{\lambda}) \\ &= \int_{\Omega^k} (\lambda_1 - \lambda_i) f_{\underline{\lambda}}(\underline{x}^*) d G(\underline{\lambda}) \end{aligned}$$

for $i = 2, 3, \dots, k$. Now

$$\int_{\Omega^k} \lambda_i f_{\underline{\lambda}}(\underline{x}^*) dG(\underline{\lambda}) = \left\{ \int_{\Omega} \lambda_i f_{\lambda_i}(x_i^*) dG_i(\lambda_i) \right\} \left\{ \prod_{\substack{j=1 \\ j \neq i}}^k \int_{\Omega} f_{\lambda_j}(x_j^*) dG_j(\lambda_j) \right\} .$$

If

$$\gamma_i(x_i^*) = \int_{\Omega} \lambda_i f_{\lambda_i}(x_i^*) dG_i(\lambda_i)$$

and

$$f_{G_j}(x_j^*) = \int_{\Omega} f_{\lambda_j}(x_j^*) dG_j(\lambda_j) ,$$

then from (4.6.3) we have for $i = 1, 2, \dots, k$,

$$(4.6.4) \quad \Delta_G(d_i, \underline{x}^*) = \{ \gamma_1(x_1^*) \prod_{j=2}^k f_{G_j}(x_j^*) \} - \{ \gamma_1(x_1^*) \prod_{\substack{j=1 \\ j \neq i}}^k f_{G_j}(x_j^*) \} .$$

If we can find functions $f_{n,j}(x_j^*)$ and $\gamma_{n,j}(x_j^*)$ of the prior observations such that

$$(4.6.5) \quad f_{n,j}(x_j^*) \xrightarrow{P} f_{G_j}(x_j^*) ,$$

and

$$(4.6.6) \quad \gamma_{n,j}(x_j^*) \xrightarrow{P} \gamma_j(x_j^*)$$

for $j = 1, 2, \dots, k$, then defining

$$(4.6.7) \quad \Delta_{i,n}(\underline{x}^*) = \{\gamma_{n,1}(\underline{x}_1^*) \prod_{j=2}^k f_{n,j}(\underline{x}_j^*)\} - \{\gamma_{n,i}(\underline{x}_i^*) \prod_{\substack{j=1 \\ j \neq i}}^k f_{n,j}(\underline{x}_j^*)\}$$

we have $\Delta_{i,n}(\underline{x}^*) \xrightarrow{P} \Delta_G(d_i, \underline{x}^*)$ for $i = 1, 2, \dots, k$, and an empirical Bayes procedure will be given by (4.3.9).

Due to the independence of the observations we have

$$(4.6.8) \quad f_{G_j}(\underline{x}_j^*) = \prod_{\ell=1}^r f_{G_j}(\underline{x}_j^{(\ell)}) = \prod_{\ell=1}^r \left\{ \int_{\Omega} f_{\lambda_j}(\underline{x}_j^{(\ell)}) dG_j(\lambda_j) \right\}.$$

The random variable X_{ij} has the same marginal density $f_{G_j}(\underline{x}_j)$ for each $i = 1, 2, \dots, n$, and therefore we can use the prior observations as well as the present observation to find an estimate of $f_{G_j}(\underline{x}_j)$ at the r points $\underline{x}_j^{(1)}, \underline{x}_j^{(2)}, \dots, \underline{x}_j^{(r)}$. Defining the empirical distribution function for the j 'th population to be

$$(4.6.9) \quad F_{n,j}(\underline{x}_j^{(\ell)}) = \frac{1}{(n+1)r} \quad (\text{total number of prior observations from the } j\text{'th population which are } \leq \underline{x}_j^{(\ell)}) ,$$

we have from (2.3.17) and (2.3.18) that

$$(4.6.10) \quad f_{n,j}(\underline{x}_j^{(\ell)}) = \frac{F_{n,j}(\underline{x}_j^{(\ell)} + h_n) - F_{n,j}(\underline{x}_j^{(\ell)} - h_n)}{2h_n} \xrightarrow{P} f_{G_j}(\underline{x}_j^{(\ell)}) ,$$

where $h_n = dn^{-(1/5)}$, $d > 0$ being some constant, for $\ell = 1, 2, \dots, r$ and $j = 1, 2, \dots, k$. Since (4.6.10) implies that

$$(4.6.11) \quad f_{n,j}(x_j^*) = \prod_{\ell=1}^r f_{n,j}(x_j^{(\ell)}) \xrightarrow{P} f_{G_j}(x_j^*)$$

for $j = 1, 2, \dots, k$, a sequence $f_{n,j}(x_j^*)$ in (4.6.7) can always be found. Therefore we are left with the problem of finding a sequence $\{\gamma_{n,j}(x_j^*)\}$ which converges in probability to $\{\gamma_j(x_j^*)\}$. If such a sequence can be found, then the empirical Bayes procedure δ_n is given by (4.3.9). Since $f_{n,j}(x_j^*) > 0$, we have from (4.6.7)

$$(4.6.12) \quad \Delta_{i,n}(\underline{x}^*) = \left\{ \frac{\gamma_{n,1}(x_1^*)}{f_{n,1}(x_1^*)} - \frac{\gamma_{n,i}(x_i^*)}{f_{n,i}(x_i^*)} \right\} f_{\pi,n}(\underline{x}^*) ,$$

where

$$(4.6.13) \quad f_{\pi,n}(\underline{x}^*) = \prod_{j=1}^k f_{n,j}(x_j^*) .$$

Now (4.6.13) is independent of i so that

$$(4.6.14) \quad \Delta_{j,n}(\underline{x}^*) = \min_{1 \leq i \leq k} \Delta_{i,n}(\underline{x}^*)$$

if and only if

$$(4.6.15) \quad \frac{\gamma_{n,j}(x_j^*)}{f_{n,j}(x_j^*)} = \max_{1 \leq i \leq k} \left\{ \frac{\gamma_{n,i}(x_i^*)}{f_{n,i}(x_i^*)} \right\} .$$

Thus the empirical Bayes procedure with respect to any $G \in \overline{\mathcal{G}}$ for selecting the best of k populations is given by

$$(4.6.16) \quad \delta_n(\underline{x}^*) = d_j \quad \text{where } j \text{ is any integer } 1, 2, \dots, k \text{ such}$$

$$\text{that } \frac{\gamma_{n,j}(\underline{x}_j^*)}{f_{n,j}(\underline{x}_j^*)} = \max_{1 \leq i \leq k} \left\{ \frac{\gamma_{n,i}(\underline{x}_i^*)}{f_{n,i}(\underline{x}_i^*)} \right\},$$

provided we can find a sequence $\{\gamma_{n,i}(\underline{x}_i^*)\}$ which converges in probability to $\{\gamma_i(\underline{x}_i^*)\}$.

As an example of the above technique, consider the case where we have the class densities given by

$$(4.6.17) \quad f_{\lambda_j}(\underline{x}_j) = \begin{cases} \lambda_j^{x_j} g(\underline{x}_j) h(\lambda_j), & x_j > a, \\ 0, & x_j \leq a, \end{cases}$$

where a is some constant and where λ_j is distributed according to G_j such that

$$G = \prod_{j=1}^k G_j \in \mathcal{G}.$$

Then

$$\begin{aligned} \gamma_j(\underline{x}_j^*) &= \int_{\Omega} \lambda_j f_{\lambda_j}(\underline{x}_j^*) dG_j(\lambda_j) \\ &= \int_{\Omega} \lambda_j \lambda_j^{\sum_{\ell=1}^r x_j^{(\ell)}} \left[\prod_{\ell=1}^r g(\underline{x}_j^{(\ell)}) \right] [h(\lambda_j)]^r dG_j(\lambda_j) \\ (4.6.18) \quad &= \frac{g(\underline{x}_j^{(1)})}{g(\underline{x}_j^{(1)} + 1)} \int_{\Omega} \left[\lambda_j^{x_j^{(1)}+1} g(\underline{x}_j^{(1)}+1) h(\lambda_j) \right] \lambda_j^{\sum_{\ell=2}^r x_j^{(\ell)}} \end{aligned}$$

$$\begin{aligned}
 & \cdot \prod_{\ell=2}^r g(x_j^{(\ell)}) (h(\lambda_j))^{r-1} dG_j(\lambda_j) \\
 &= \frac{g(x_j^{(1)})}{g(x_j^{(1)}+1)} \int_{\Omega} f_{\lambda_j}(x_j^{(1)}+1) \prod_{\ell=2}^r f_{\lambda_j}(x_j^{(\ell)}) dG_j(\lambda_j) \\
 &= \frac{g(x_j^{(1)})}{g(x_j^{(1)}+1)} f_{G_j}(x_j^{(1)}+1) \prod_{\ell=2}^r f_{G_j}(x_j^{(\ell)}) \\
 &= \frac{g(x_j^{(1)})}{g(x_j^{(1)}+1)} \cdot \frac{f_{G_j}(x_j^{(1)}+1)}{f_{G_j}(x_j^{(1)})} \cdot f_{G_j}(x_j^*),
 \end{aligned}$$

where $x_j^{(1)}$ is arbitrary and determined by convenience only, and where

$$f_{G_j}(x_j^*) = \prod_{\ell=1}^r f_{G_j}(x_j^{(\ell)}) = \int_{\Omega} \prod_{\ell=1}^r f_{\lambda_j}(x_j^{(\ell)}) dG_j(\lambda_j) .$$

Since the function g is known, and $f_{n,j}(x_j^{(\ell)})$ defined in (4.6.10) converges in probability to $f_{G_j}(x_j^{(\ell)})$, then $\gamma_{n,j}(x_j^*)$ defined by

$$(4.6.19) \quad \gamma_{n,j}(x_j^*) = \frac{g(x_j^{(1)})}{g(x_j^{(1)}+1)} f_{n,j}(x_j^{(1)}+1) \prod_{\ell=2}^r f_{n,j}(x_j^{(\ell)})$$

converges in probability to $\gamma_j(x_j^*)$. Writing

$$(4.6.20) \quad \gamma_{n,j}(x_j^*) = \left\{ \frac{g(x_j^{(1)})}{g(x_j^{(1)}+1)} \frac{f_{n,j}(x_j^{(1)}+1)}{f_{n,j}(x_j^{(1)})} \right\} f_{n,j}(x_j^*) ,$$

where

$$f_{n,j}(x_j^*) = \prod_{\ell=1}^r f_{n,j}(x_j^{(\ell)}) ,$$

and since $f_{n,j}(x_j^{(1)}) > 0$, we see that the empirical Bayes procedure for selecting the best of k populations with respect to any $G \in \bar{\mathcal{J}}$ and using loss function (4.2.2) is given by

$$(4.6.21) \quad \delta_n(\underline{x}^*) = d_j \text{ where } j \text{ is any integer } 1, 2, \dots, k \text{ such that}$$

$$\frac{g(x_j^{(1)})}{g(x_j^{(1)}+1)} \frac{f_{n,j}(x_j^{(1)}+1)}{f_{n,j}(x_j^{(1)})} = \max_{1 \leq i \leq k} \left\{ \frac{g(x_i^{(1)})}{g(x_i^{(1)}+1)} \frac{f_{n,i}(x_i^{(1)}+1)}{f_{n,i}(x_i^{(1)})} \right\}.$$

After a suitable transformation a large class of densities can be written in the form (4.6.17).

CHAPTER V

NON-PARAMETRIC EMPIRICAL BAYES ESTIMATION

AND HYPOTHESIS TESTING

The non-parametric case occurs when the class of conditional probability distributions of X given λ is not restricted to a particular parametric family but is a much larger class of probability distributions which cannot be defined in terms of a finite number of parameters. In the discrete case, for example, it may be the class of all probability functions assigning probability one to some specified denumerable set of numbers $\chi = \{x\}$. Let $\mathcal{F}_1 = \{F_\omega(x) : \omega \in \bar{\Omega}\}$, the class of all conditional distribution functions such that probability one is assigned to χ where $\bar{\Omega} = \{\omega\}$ is an abstract indexing set. Let μ be the a priori probability measure defined on \mathcal{A} , a σ -algebra of subsets of $\bar{\Omega}$, and let Y be an $\bar{\Omega}$ -valued random variable which is the identity mapping of $\bar{\Omega}$ onto itself.

5.1. Estimation: Discrete Case.

Consider the random variables X_1, X_2, \dots, X_r which are conditionally independent and identically distributed with common conditional distribution function $F_\omega(x)$ given that $Y = \omega$. Define the random variable Λ for a given measurable function $h(x)$ by

$$(5.1.1) \quad \Lambda = \Lambda(Y) = E(h(X) \mid Y),$$

where X is a generic representation of the X_j 's , and assume that the a priori probability measure space $(\bar{\Omega}, \mathcal{A}, \mu)$ is such that

$$(5.1.2) \quad E h^2(X) < \infty .$$

Using $\underline{X} = (X_1, X_2, \dots, X_r)$, the vector of observations, we wish to obtain an estimate of the value λ of Λ . If $\psi(\underline{x})$ is an estimator of Λ , we will use the usual squared error loss function

$$(5.1.3) \quad L(\psi(\underline{x}), \lambda) = (\psi(\underline{x}) - \lambda)^2 ,$$

where $\underline{x} = (x_1, x_2, \dots, x_r)$. The risk involved in using any estimator $\psi(\underline{x})$ is

$$(5.1.4) \quad R(\psi) = EL(\psi(\underline{X}), \Lambda) = E[\psi(\underline{X}) - \Lambda]^2 .$$

From (5.1.1) and (5.1.2) we see that

$$(5.1.5) \quad \begin{aligned} E\Lambda^2 &= E\{E^2(h(X) | Y)\} \leq E\{E(h^2(X) | Y)\} \\ &= Eh^2(X) < \infty , \end{aligned}$$

and hence

$$(5.1.6) \quad \begin{aligned} R(\psi) &= E\{E(\psi(\underline{X}) - \Lambda)^2 \mid \underline{X} = \underline{x}\} \\ &= E\{[\psi^2(\underline{X}) - 2\psi(\underline{X})E(\Lambda | \underline{X}) + E^2(\Lambda | \underline{X})] + E(\Lambda^2 | \underline{X}) - E^2(\Lambda | \underline{X})\} \\ &= E\{[\psi(\underline{X}) - E(\Lambda | \underline{X})]^2 + E(\Lambda^2 | \underline{X}) - E^2(\Lambda | \underline{X})\} \\ &= E\{[\psi(\underline{X}) - E(\Lambda | \underline{X})]^2\} + E\{\text{Var}(\Lambda | \underline{X})\} . \end{aligned}$$

This is a minimum when $\psi(\underline{X}) = E(\Lambda | \underline{X})$ and hence the Bayes estimator

$\psi_{\mu}(\underline{x})$ which minimizes $R(\psi)$ is

$$(5.1.7) \quad \psi_{\mu}(\underline{x}) = E(\Lambda | \underline{X} = \underline{x})$$

for all $\underline{x} \in \chi^* = \{\underline{x} : P(\underline{X} = \underline{x}) > 0\}$, and from (5.1.6) the risk of the Bayes estimator is

$$(5.1.8) \quad R(\psi_{\mu}) = E\Lambda^2 - E\psi_{\mu}^2(\underline{X}) < \infty.$$

As before the a priori probability structure of the problem must be known before we can obtain the Bayes estimator in (5.1.7). Assume that this structure is unknown but we have additional information in the form of vectors of observations

$$\underline{X}_i = (X_{i,1}, X_{i,2}, \dots, X_{i,r+1}) \quad \text{for } i = 1, 2, \dots, n$$

where the \underline{X}_i 's are mutually independent and independent of \underline{X} and where for each i the $X_{i,j}$'s are conditionally independent and identically distributed according to $F_{\omega_i}(x)$ given that $Y_i = \omega_i$ where Y_i , $i = 1, 2, \dots, n$ are mutually independent of Y and have the same distribution as Y .

If we let $\underline{X}_i^{(r)} = (X_{i,1}, X_{i,2}, \dots, X_{i,r})$, then $\underline{X}_i^{(r)}$ and $E(h(X_{i,r+1}) | Y_i)$ have the same joint distribution as \underline{X} and Λ so that

$$\begin{aligned}
 (5.1.9) \quad & E[h(X_{i,r+1}) | \underline{X}_i^{(r)} = \underline{x}] \\
 &= E\{E[h(X_{i,r+1}) | Y_i, \underline{X}_i^{(r)}] | \underline{X}_i^{(r)} = \underline{x}\} \\
 &= E\{E[h(X_{i,r+1}) | Y_i] | \underline{X}_i^{(r)} = \underline{x}\} \\
 &= E(\Lambda | \underline{X} = \underline{x}) = \psi_\mu(\underline{x}) .
 \end{aligned}$$

This suggests the following empirical Bayes estimation procedure:

If $\underline{x} = (x_1, x_2, \dots, x_r)$, then for each \underline{x} we can permute the components to get $m(\underline{x})$ distinct vectors with $1 \leq m(\underline{x}) \leq r!$. If $\underline{x}_{(q)}$, $q = 1, 2, \dots, m(\underline{x})$ denote the distinct vectors, then define the random functions $M_i(\underline{x})$, $i = 1, 2, \dots, n$ and $\bar{M}_n(\underline{x})$ by

$$(5.1.10) \quad M_i(\underline{x}) = \begin{cases} 1, & \text{if there exists a } q, 1 \leq q \leq m(\underline{x}), \text{ such that } \underline{X}_i^{(r)} = \underline{x}_{(q)} , \\ 0, & \text{otherwise ,} \end{cases}$$

and

$$(5.1.11) \quad \bar{M}_n(\underline{x}) = \sum_{i=1}^n M_i(\underline{x}) .$$

If the empirical Bayes estimator is defined by

$$(5.1.12) \quad \psi_n(\underline{x}) = \begin{cases} \frac{1}{\bar{M}_n(\underline{x})} \sum_{i=1}^n M_i(\underline{x}) h(x_{i,r+1}) , & \bar{M}_n(\underline{x}) > 0 , \\ 0 & , \text{ otherwise ,} \end{cases}$$

then Johns [6] proves

Theorem 6: Let X be a generic representation of $r \geq 1$ independent random variables which are identically distributed according to $F_{\omega}(x) \in \mathcal{F}_1$, and let there be a measurable function $h(x)$ such that (5.1.1) holds. Let X be the vector of the present r observations, and let \underline{X}_i , $i = 1, 2, \dots, n$ be the prior observations where the \underline{X}_i 's are mutually independent and independent of X and where each \underline{X}_i consists of $r+1$ independent random variables which are identically distributed according to $F_{\omega_i}(x) \in \mathcal{F}_1$. If $(\bar{\Omega}, \mathcal{Q}, \mu)$ is such that (5.1.2) holds, then, using loss function (5.1.3),

$$(5.1.13) \quad \lim_{n \rightarrow \infty} R(\psi_n) = R(\psi_{\mu}) ,$$

where $R(\psi_{\mu})$ is the risk of the Bayes estimator ψ_{μ} and $R(\psi_n)$ is the risk of the empirical Bayes estimator ψ_n defined in (5.1.12).

Instead of proving the above theorem, we will prove, in the following section, the above result for a slightly different case; namely the supplementary sample method of Krutchkoff [8] which was derived from Johns' non-parametric result by dropping the requirement that an unbiased estimate of λ be known at the time the estimate is to be made. The reader is referred to [6] for Johns' original proof.

5.2. Supplementary Sample Method.

Suppose that the random variable X has an unknown conditional distribution function $F_{\lambda}(x)$ given $\Lambda = \lambda$ which may be specified by an unknown probability mass function $P_{\lambda}(x)$ (i.e. X is a discrete

random variable). One is required to use the observed value of X to obtain an estimate of λ which has distribution $G(\lambda)$, a specific but unknown member of

$$(5.2.1) \quad \mathcal{G} = \{G(\lambda) : E\Lambda^2 = \int \lambda^2 d G(\lambda) < \infty\} .$$

Suppose that in our replications we have a supplementary sample y which is the realization of the random variable Y with conditional distribution function $H_\lambda(y)$ given $\Lambda = \lambda$, a specific but unknown member of

$$(5.2.2) \quad \mathcal{H} = \{H_\lambda(y) : E_\lambda Y = \int y d H_\lambda(y) = \lambda, \text{ and} \\ EY^2 = \int y^2 d H_\lambda(y) d G(\lambda) < \infty\} .$$

Thus the expectation of Y is the realization of Λ for that replication and not necessarily the present value of λ . Thus after an action has been taken we can observe with error, perhaps much later, the value of λ . For example, we may observe later the consequences of the use of some product.

Assume that X and Y are conditionally independent and that the replications of the problem are independent and identically distributed. Then the joint distribution of (X, Y) for $\Lambda = \lambda$ is given by

$$(5.2.3) \quad F_\lambda(x) H_\lambda(y) ,$$

and if (Λ_i, X_i, Y_i) represents the i 'th replication, then

$(\Lambda_1, X_1, Y_1, \Lambda_2, X_2, Y_2, \dots, \Lambda_n, X_n, Y_n)$ has the distribution

$$(5.2.4) \quad \prod_{i=1}^n G(\lambda_i) F_{\lambda_i}(x_i) H_{\lambda_i}(y_i) .$$

From (5.1.6) with $r = 1$, we saw that the estimate of λ which minimizes the expected squared error $R(\psi) = E(\psi(X) - \Lambda)^2$, where $\psi(X)$ is an estimator of Λ , is

$$(5.2.5) \quad \psi_{\mu}(x) = E(\Lambda \mid X = x) ,$$

and from (5.1.8) the Bayes risk is

$$(5.2.6) \quad R(\psi_{\mu}) = E\{\text{Var}(\Lambda \mid X)\} < \infty .$$

The "regret" is defined to be

$$(5.2.7) \quad r(\psi) = R(\psi) - R(\psi_{\mu}) = E[\psi(X) - E(\Lambda \mid X)]^2 .$$

Assuming that $E(\Lambda \mid X)$ is not known to us, we wish to use the $2n$ -tuple of values $\tilde{z}_n = (x_1, y_1, x_2, y_2, \dots, x_n, y_n)$, the realization of

$$(5.2.8) \quad \tilde{z}_n = (X_1, Y_1, X_2, Y_2, \dots, X_n, Y_n)$$

for n previous occurrences of the problem in order to obtain an estimate

$\psi_n(\tilde{z}_n; x)$ of λ . Note here that the expectations must be taken over

\tilde{z}_n as well as X and Λ from now on since $r(\psi)$ is an unconditional

and $\phi_1, \phi_2, \dots, \phi_n$ are the solutions of the system of equations

$$\phi_1^2 + \phi_2^2 + \dots + \phi_n^2 = 1 \quad (2.1)$$

and $\phi_1, \phi_2, \dots, \phi_n$ are the solutions of the system of equations

$$\phi_1^2 + \phi_2^2 + \dots + \phi_n^2 = 1 \quad (2.2)$$

and $\phi_1, \phi_2, \dots, \phi_n$ are the solutions of the system of equations

$$\phi_1^2 + \phi_2^2 + \dots + \phi_n^2 = 1 \quad (2.3)$$

and $\phi_1, \phi_2, \dots, \phi_n$ are the solutions of the system of equations

$$\phi_1^2 + \phi_2^2 + \dots + \phi_n^2 = 1 \quad (2.4)$$

and $\phi_1, \phi_2, \dots, \phi_n$ are the solutions of the system of equations

$$\phi_1^2 + \phi_2^2 + \dots + \phi_n^2 = 1 \quad (2.5)$$

and $\phi_1, \phi_2, \dots, \phi_n$ are the solutions of the system of equations

$$\phi_1^2 + \phi_2^2 + \dots + \phi_n^2 = 1 \quad (2.6)$$

and $\phi_1, \phi_2, \dots, \phi_n$ are the solutions of the system of equations

expectation. X and Λ are independent of Z_n by the assumption above.

Let

$$(5.2.9) \quad \delta(x_i; x) = \begin{cases} 1, & \text{if } x_i = x, \\ 0, & \text{if } x_i \neq x, \end{cases}$$

and

$$(5.2.10) \quad m_n(z_n; x) = \sum_{i=1}^n \delta(x_i, x) .$$

Define

$$(5.2.11) \quad \psi_n(z_n; x) = \begin{cases} \frac{1}{m_n(z_n; x)} \sum_{i=1}^n \delta(x_i, x) y_i, & \text{if } m_n(z_n; x) > 0 . \\ 0, & \text{if } m_n(z_n; x) = 0 . \end{cases}$$

Now

$$\begin{aligned} E(\Lambda | x) &= E\{E(Y | \Lambda) | X = x\} \quad \text{a.s.} \\ (5.2.12) \quad &= E\{E(Y | \Lambda, X) | X = x\} \quad \text{a.s.} \\ &= E(Y | x) \quad \text{a.s.} \end{aligned}$$

since $E(Y | \Lambda) = \Lambda$ from (5.2.2), and X and Y are conditionally independent by assumption. In the third step we integrate with respect to Λ only.

When $m_n(z_n; x) = m > 0$, $\psi_n(z_n; x)$ is the average of m independent unbiased estimates of $E(Y | x) = E(\Lambda | x)$. Therefore

$$(5.2.13) \quad E[\psi_n(Z_n; x) | m_n = m > 0, X = x] = \psi_\mu(x) ,$$

$$(5.2.14) \quad E[\psi_n(Z_n; x) | m_n = 0, X = x] = 0 ,$$

$$(5.2.15) \quad E[(\psi_n(Z_n; x) - \psi_\mu(x))^2 | m_n = m > 0, X = x] = \frac{1}{m} \text{Var} (Y|x) ,$$

and

$$(5.2.16) \quad E[(\psi_n(Z_n; x) - \psi_\mu(x))^2 | m_n = 0, X = x] = \psi_\mu^2(x)$$

If we replace ψ by ψ_n in (5.2.7), we see that

$$\begin{aligned} r(\psi_n) &= E[\psi_n(Z_n; X) - \psi_\mu(X)]^2 \\ &= E\{E[(\psi_n(Z_n; X) - \psi_\mu(X))^2 | X = x]\} \\ &= E\{E[E[(\psi_n(Z_n; X) - \psi_\mu(X))^2 | m_n, X] | X = x]\} \\ (5.2.17) \quad &= E\{E\left[\sum_{m=1}^n \frac{1}{m} \text{Var} (Y|X) \text{Prob} (m_n(Z_n; X) = m) \right. \\ &\quad \left. + \psi_\mu^2(X) \text{Prob} (m_n(Z_n; X) = 0) | X = x]\right\} \\ &= E\{\text{Var} (Y|X)E\left[\sum_{m=1}^n \frac{1}{m} \text{Prob} (m_n(Z_n; X) = m) | X = x\right] \\ &\quad + \psi_\mu^2(X)E[\text{Prob} (m_n(Z_n; X) = 0) | X = x]\} \end{aligned}$$

Also,

$$\begin{aligned}
 r(\psi_n) &\leq E\{\text{Var}(Y|x) + \psi_\mu^2(x)\} \\
 (5.2.18) \quad &= E\{E(Y^2|x) - E^2(Y|x) + E^2(\Lambda|x)\} \\
 &= E\{E(Y^2|x)\} < \infty
 \end{aligned}$$

because of (5.2.2).

Now for any Z_n such that $m_n(Z_n; x) = s$,

$$(5.2.19) \quad \text{Prob}(m_n(Z_n; x) = s) = \binom{n}{s} P^s(x) (1-P(x))^{n-s},$$

where

$$(5.2.20) \quad P(x) = \int P_\lambda(x) dG(\lambda)$$

is the probability that there are exactly s successes in n independent trials with probability $P(x)$ for success. Then

$$(5.2.21) \quad E[\text{Prob}(m_n(Z_n; X) = 0) | X = x] = (1 - P(x))^n,$$

and hence

$$(5.2.22) \quad \lim_{n \rightarrow \infty} E[\text{Prob}(m_n(Z_n; X) = 0) | X = x] = 0 \text{ a.s.}$$

Also,

$$\begin{aligned}
 (5.2.23) \quad E\left[\sum_{m=1}^n \frac{1}{m} \text{Prob}(m_n(Z_n; X) = m) | X = x\right] \\
 = \sum_{m=1}^n \frac{1}{m} \binom{n}{m} P^m(x) (1 - P(x))^{n-m}
 \end{aligned}$$

$$\begin{aligned}
 &\leq \sum_{m=1}^n \frac{2}{m+1} \binom{n}{m} P^m(x) (1 - P(x))^{n-m} \\
 &\leq \frac{2}{(n+1)P(x)} \sum_{m=1}^n \binom{n+1}{m+1} P^{m+1}(x) (1 - P(x))^{n-m} \\
 &\leq \frac{2}{(n+1)P(x)} .
 \end{aligned}$$

Then

$$(5.2.24) \quad \lim_{n \rightarrow \infty} E \left[\sum_{m=1}^n \frac{1}{m} \text{Prob} (m_{\tilde{Z}_n; X} = m) | X = x \right] = 0 \quad \text{a.s.} ,$$

and by (5.2.17), (5.2.18), (5.2.22), (5.2.24), and the Lebesgue dominated convergence theorem, we see that

$$(5.2.25) \quad \lim_{n \rightarrow \infty} r(\psi_n) = 0 .$$

Thus the risk of $\psi_{\tilde{Z}_n; x}$ attains the risk of the Bayes estimator as $n \rightarrow \infty$, and hence $\psi_{\tilde{Z}_n; x}$ is an asymptotically optimal estimator.

5.3. Estimation: Continuous Case.

Johns [6] also considers the case where the observed X 's possess absolutely continuous distribution functions. Let

$\mathcal{F}_2 = \{F_\omega(x) : \omega \in \bar{\Omega}\}$ be the class of all absolutely continuous conditional distribution functions where $\bar{\Omega} = \{\omega\}$ is an abstract indexing set. If $(\bar{\Omega}, \mathcal{A}, \mu)$ is an a priori probability measure space where

\mathcal{A} is a σ -algebra of subsets of $\bar{\Omega}$, then there exists a function $f_{\omega}(u)$ defined on the product space (reals) $\times \bar{\Omega}$ such that

$$(5.3.1) \quad F_{\omega}(x) = \int_{-\infty}^x f_{\omega}(u) du, \quad ,$$

for each $\omega \in \bar{\Omega}$. Assume that $(\bar{\Omega}, \mathcal{A})$ is such that the function $f_{\omega}(u)$ is a measurable function on the product space (reals) $\times \bar{\Omega}$. As before $Y = Y(\omega)$ is the $\bar{\Omega}$ -valued random variable which is the identity mapping of $\bar{\Omega}$ onto itself. Let $\underline{X} = (X_1, X_2, \dots, X_r)$ where X_j , $j = 1, 2, \dots, r$ are random variables which are conditionally independent and identically distributed according to $F_{\omega}(x)$ given that $Y = \omega$.

Define the random variable Λ for a given measurable function $h(x)$ by

$$(5.3.2) \quad \Lambda = \Lambda(Y) = E(h(X) | Y), \quad ,$$

where X is a generic representation of the X_j 's and assume

$$(5.3.3) \quad E h^2(X) < \infty \quad .$$

As in the discrete case, the Bayes estimator of Λ using \underline{X} where the risk is the expected squared error is given by

$$(5.3.4) \quad \psi_{\mu}(\underline{x}) = E(\Lambda | \underline{X} = \underline{x}) \quad .$$

As before we have the random vectors of prior observations $\underline{X}_i = (X_{i,1}, X_{i,2}, \dots, X_{i,r+1})$ for $i = 1, 2, \dots, n$, the \underline{X}_i 's being

independent of each other and of \underline{X} and where for each i the $X_{i,j}$'s are conditionally independent and identically distributed according to $F_{\omega_i}(x)$ given that $Y_i = \omega_i$ where Y_i , $i = 1, 2, \dots, n$ are independent random variables and independent of Y such that each has the same distribution as Y . If $\underline{X}_i^{(r)} = (X_{i,1}, X_{i,2}, \dots, X_{i,r})$, $i = 1, 2, \dots, n$, then, as before

$$(5.3.5) \quad E(h(X_{i,r+1}) | \underline{X}_i^{(r)} = \underline{x}) = E(\Lambda | \underline{X} = \underline{x}) = \psi_{\mu}(\underline{x}) .$$

The X 's can be made discrete in the following way. Consider the double sequence of half-open intervals

$$(5.3.6) \quad I_t^{(n)} = \left[\frac{tc}{n^{1-\delta/r}}, \frac{(t+1)c}{n^{1-\delta/r}} \right), \quad t = 0, \pm 1, \pm 2, \dots; n = 1, 2, \dots ,$$

where $c > 0$ and $0 < \delta < 1$. For each n partition r -dimensional Euclidean space into a countable sequence of non-overlapping hypercubes $G_j^{(n)}$ where

$$(5.3.7) \quad G_j^{(n)} = I_{t_{1,j}}^{(n)} \times I_{t_{2,j}}^{(n)} \times \dots \times I_{t_{r,j}}^{(n)}, \quad j = 1, 2, \dots ,$$

where the $t_{i,j}$'s are suitably chosen integers. For each n let $G_j^{(n)}(\underline{x})$ be the unique member of the sequence (5.3.7) which contains the r -component numerical vector $\underline{x} = (x_1, x_2, \dots, x_r)$. If $\underline{x}_{(q)}$, $q = 1, 2, \dots, m(\underline{x})$, $m(\underline{x}) \geq 1$, are the distinct vectors obtained by permuting the components of \underline{x} , we have, analogous to (5.1.10) and (5.1.11),

$$(5.3.8) \quad M_i^{(n)}(\underline{x}) = \begin{cases} 1, & \text{if there exists a } q, 1 \leq q \leq m(\underline{x}) \text{ such that } X_i^{(r)} \in C^{(n)}(\underline{x}_{(q)}) \\ 0, & \end{cases}$$

for $i = 1, 2, \dots, n$, and

$$(5.3.9) \quad \bar{M}^{(n)}(\underline{x}) = \sum_{i=1}^n M_i^{(n)}(\underline{x}) .$$

If we define the empirical Bayes estimator $\psi_n(\underline{x})$ by

$$(5.3.10) \quad \psi_n(\underline{x}) = \begin{cases} \frac{1}{\bar{M}^{(n)}(\underline{x})} \sum_{i=1}^n M_i^{(n)}(\underline{x}) h(x_{i,r+1}) , & \bar{M}^{(n)}(\underline{x}) > 0 , \\ 0 & , \text{ otherwise ,} \end{cases}$$

then Johns [6] proves

Theorem 7: Let X be a generic representation of $r \geq 1$ random variables which are independent and identically distributed according to $F_\omega \in \mathcal{F}_2$, and let there be a measurable function $h(x)$ such that (5.3.2) holds. Suppose $(\bar{\Omega}, \mathcal{A}, \mu)$ is such that there exists a measurable function f_ω such that (5.3.1) holds. Let \underline{X} be the vector of the present r observations, and let \underline{X}_i , $i = 1, 2, \dots, n$ be the prior observations where the \underline{X}_i 's are mutually independent and independent of \underline{X} and where each \underline{X}_i consists of $r+1$ independent random variables which are identically distributed according to $F_{\omega_i} \in \mathcal{F}_2$. If $(\bar{\Omega}, \mathcal{A}, \mu)$ is such that (5.3.3) holds, then, using loss function (5.1.3),

$$(5.3.11) \quad \lim_{n \rightarrow \infty} R(\psi_n) = R(\psi_\mu) ,$$

where $R(\psi_\mu)$ is the risk of the Bayes estimator ψ_μ and $R(\psi_n)$ is the risk of the empirical Bayes estimator ψ_n defined in (5.3.10).

5.4. Application to Hypothesis Testing.

The empirical Bayes estimation procedures just described can be applied to two-decision problems of the hypothesis testing type.

Suppose we wish to test a hypothesis of the type (2.3.12); i.e.,

$H_0 : \lambda \leq \lambda^*$, with loss function (2.3.11); i.e.

$$L(d_0, \lambda) = \begin{cases} 0 & , \text{ if } \lambda \leq \lambda^* , \\ (\lambda - \lambda^*) & , \text{ if } \lambda > \lambda^* , \end{cases}$$

$$L(d_1, \lambda) = \begin{cases} (\lambda^* - \lambda) & , \text{ if } \lambda \leq \lambda^* , \\ 0 & , \text{ if } \lambda > \lambda^* , \end{cases}$$

where λ^* is a fixed constant.

If $\psi(\underline{x})$ is the function which assigns a decision $\psi(\underline{x}) = d_0$ or d_1 to each possible value $\underline{x} = (x_1, x_2, \dots, x_r)$ of the random variable $\underline{X} = (X_1, X_2, \dots, X_r)$, then let..

$$(5.4.1) \quad \delta(\underline{x}) = \begin{cases} 0 & , \text{ when } \psi(\underline{x}) = d_0 . \\ 1 & , \text{ when } \psi(\underline{x}) = d_1 . \end{cases}$$

Then

$$\begin{aligned}
 L(\psi(\underline{x}), \lambda) &= \delta(\underline{x}) L(d_1, \lambda) + (1 - \delta(\underline{x})) L(d_0, \lambda) \\
 (5.4.2) \quad &= L(d_0, \lambda) - \delta(\underline{x}) [L(d_0, \lambda) - L(d_1, \lambda)] \\
 &= L(d_0, \lambda) - \delta(\underline{x}) [\lambda - \lambda^*]
 \end{aligned}$$

The risk is then

$$\begin{aligned}
 R(\psi) &= EL(\psi(\underline{X}), \Lambda) \\
 &= EL(d_0, \Lambda) - E[\delta(\underline{X})(\Lambda - \lambda^*)] \\
 (5.4.3) \quad &= EL(d_0, \Lambda) - E\{\delta(\underline{X}) E[(\Lambda - \lambda^*) | \underline{X}]\} \\
 &= EL(d_0, \Lambda) - E\{\delta(\underline{X}) [E(\Lambda | \underline{X}) - \lambda^*]\} .
 \end{aligned}$$

The function $\psi_\mu(\underline{x})$ which minimizes $R(\psi)$ is the $\psi_\mu(\underline{x})$ corresponding to

$$(5.4.4) \quad \delta_\mu(\underline{x}) = \begin{cases} 0, & \text{if } \psi_\mu(\underline{x}) = E(\Lambda | \underline{X}) \leq \lambda^* . \\ 1, & \text{if } \psi_\mu(\underline{x}) = E(\Lambda | \underline{X}) > \lambda^* . \end{cases}$$

If we consider the supplementary sample approach, then $r = 1$ above, and the Bayes risk is then

$$\begin{aligned}
 (5.4.5) \quad R(\psi_\mu) &\leq EL(d_0, \Lambda) \\
 &\leq E|\Lambda - \lambda^*| < \infty
 \end{aligned}$$

due to (5.2.1). The regret in using arbitrary $\psi(\underline{x})$ with corresponding $\delta(\underline{x})$ is

$$\begin{aligned}
 (5.4.6) \quad r(\psi) &= R(\psi) - R(\psi_\mu) \\
 &= E\{(\delta_\mu(X) - \delta(X))(\psi_\mu(X) - \lambda^*)\} .
 \end{aligned}$$

Using our empirical Bayes estimate $\psi_n(z_n; x)$ defined in (5.2.11), we define a supplementary sample non-parametric empirical Bayes test of the hypothesis as the $\psi_n(z_n; x)$ such that

$$(5.4.7) \quad \delta_n(z_n; x) = \begin{cases} 0 , & \text{if } \psi_n(z_n; x) \leq \lambda^* . \\ 1 , & \text{if } \psi_n(z_n; x) > \lambda^* . \end{cases}$$

The regret becomes

$$(5.4.8) \quad r(\psi_n) = E\{(\delta_\mu(X) - \delta_n(z_n; X))(\psi_\mu(X) - \lambda^*)\} .$$

Now

$$(5.4.9) \quad \delta_\mu(x) - \delta_n(z_n; x) = 1 \quad \text{only when } \psi_\mu(x) - \lambda^* > 0 ,$$

and

$$(5.4.10) \quad \delta_\mu(x) - \delta_n(z_n; x) = -1 \quad \text{only when } \psi_\mu(x) - \lambda^* \leq 0 .$$

Also,

$$\begin{aligned}
 (5.4.11) \quad \delta_\mu(x) - \delta_n(z_n; x) &\neq 0 \quad \text{only when } \psi_\mu(x) - \lambda^* > 0 \\
 &\quad \text{and } \psi_n(z_n; x) - \lambda^* \leq 0 , \\
 &\quad \text{or } \psi_\mu(x) - \lambda^* \leq 0 \\
 &\quad \text{and } \psi_n(z_n; x) - \lambda^* > 0 .
 \end{aligned}$$

Then (5.4.8), (5.4.9), (5.4.10), and (5.4.11) imply that

$$\begin{aligned}
 (5.4.12) \quad r(\psi_n) &\leq E |\psi_\mu(X) - \lambda^*| \\
 &\leq E |(\psi_\mu(X) - \lambda^*) - (\psi_n(Z_n; X) - \lambda^*)| \\
 &\leq E |\psi_n(Z_n; X) - \psi_\mu(X)| .
 \end{aligned}$$

From (5.2.7) and (5.2.25) we have proved that

$$(5.4.13) \quad \lim_{n \rightarrow \infty} E(\psi_n(Z_n; X) - \psi_\mu(X))^2 = 0$$

which implies that

$$(5.4.14) \quad \lim_{n \rightarrow \infty} r(\psi_n) = 0 .$$

Thus $\lim_{n \rightarrow \infty} R(\psi_n) = R(\psi_\mu)$, and hence $\psi_n(z_n; x)$ defined by (5.2.11)

such that (5.4.7) holds is an asymptotically optimal test of the hypothesis.

In a similar manner Johns [6] showed originally that if

$$(5.4.15) \quad \delta_n(\underline{x}) = \begin{cases} 0, & \text{if } \psi_n(\underline{x}) \leq \lambda^* , \\ 1, & \text{if } \psi_n(\underline{x}) > \lambda^* , \end{cases}$$

where $\psi_n(\underline{x})$ is defined by (5.1.12), then $\lim_{n \rightarrow \infty} R(\delta_n) = R(\delta_\mu)$ whenever the a priori probability measure is such that $\psi_n(\underline{x}) \xrightarrow{P} E(\Lambda | X = \underline{x})$ for all \underline{x} in some set S which is assigned probability one under the distribution of \underline{X} , and $E|\Lambda| < \infty$. Thus it is clear how to obtain an asymptotically optimal test of the hypothesis in the case where an

unbiased estimate of λ in the distribution of X is known at the time the estimate is to be made.

For this case Johns also shows how to obtain a test of a hypothesis of the type (2.3.13); i.e. $H_0 : |\lambda - \lambda^*| \leq \Delta$, where λ^* and $\Delta > 0$ are fixed, and with loss function (2.3.14); i.e.

$$L(d_0, \lambda) = \begin{cases} (\lambda - \lambda^*)^2 - \Delta^2, & \text{if } |\lambda - \lambda^*| > \Delta, \\ 0, & \text{if } |\lambda - \lambda^*| \leq \Delta, \end{cases}$$

$$L(d_1, \lambda) = \begin{cases} 0, & \text{if } |\lambda - \lambda^*| > \Delta, \\ \Delta^2 - (\lambda - \lambda^*)^2, & \text{if } |\lambda - \lambda^*| \leq \Delta. \end{cases}$$

If $\delta(\underline{x})$ is defined by (5.4.1), we see, as before, that the risk is

$$\begin{aligned} R(\psi) &= EL(\psi(\underline{X}), \Lambda) \\ (5.4.16) \quad &= EL(d_0, \Lambda) - E[\delta(\underline{X}) ((\Lambda - \lambda^*)^2 - \Delta^2)] \\ &= EL(d_0, \Lambda) - E\{\delta(\underline{X}) [E(\Lambda^2 | \underline{X}) - 2\lambda^* E(\Lambda | \underline{X}) + \lambda^{*2} - \Delta^2]\}. \end{aligned}$$

Therefore the function $\psi_\mu(\underline{x})$ which minimizes $R(\psi)$ is the $\psi_\mu(\underline{x})$ corresponding to

$$(5.4.17) \quad \delta_\mu^*(\underline{x}) = \begin{cases} 0, & \text{if } \psi_\mu(\underline{x}) = E(\Lambda^2 | \underline{X} = \underline{x}) - 2\lambda^* E(\Lambda | \underline{X} = \underline{x}) \leq \Delta^2 - \lambda^{*2} \\ 1, & \text{if } \psi_\mu(\underline{x}) = E(\Lambda^2 | \underline{X} = \underline{x}) - 2\lambda^* E(\Lambda | \underline{X} = \underline{x}) > \Delta^2 - \lambda^{*2} \end{cases}$$

Johns then shows that if we can find empirical Bayes estimates $\psi_n^{(1)}(\underline{x})$ and $\psi_n^{(2)}(\underline{x})$ based on the prior independent observations such that

$$(5.4.18) \quad \psi_n^{(1)}(\underline{x}) \xrightarrow{P} E(\Lambda | \underline{X} = \underline{x}) ,$$

and

$$(5.4.19) \quad \psi_n^{(2)}(\underline{x}) \xrightarrow{P} E(\Lambda^2 | \underline{X} = \underline{x}) ,$$

for all $\underline{x} \in S$, and if $E\Lambda^2 < \infty$ and we define

$$(5.4.20) \quad \delta_n^*(\underline{x}) = \begin{cases} 0 , & \text{if } \psi_n(\underline{x}) = \psi_n^{(2)}(\underline{x}) - 2\lambda^* \psi_n^{(1)}(\underline{x}) \leq \Delta^2 - \lambda^{*2} , \\ 1 , & \text{if } \psi_n(\underline{x}) = \psi_n^{(2)}(\underline{x}) - 2\lambda^* \psi_n^{(1)}(\underline{x}) > \Delta^2 - \lambda^{*2} , \end{cases}$$

then

$$\lim_{n \rightarrow \infty} R(\delta_n^*) \longrightarrow R(\delta_\mu^*) .$$

An empirical Bayes estimate $\psi_n^{(1)}(\underline{x})$ for $E(\Lambda | \underline{X} = \underline{x})$ is given by (5.1.12). To get an empirical Bayes estimate of $E(\Lambda^2 | \underline{X} = \underline{x})$, we assume that $Eh^4(X) < \infty$ and that the number of components in the vector of prior observations \underline{X}_i for $i = 1, 2, \dots, n$ exceeds the number in \underline{X} by at least two. It can then be shown that

$$(5.4.21) \quad E(h(X_{1,r+1}) h(X_{1,r+2}) | \underline{X}_1^{(r)} = \underline{x}) = E(\Lambda^2 | \underline{X} = \underline{x})$$

and that

$$(5.4.22) \quad \psi_n^{(2)}(\underline{x}) = \begin{cases} \frac{1}{\bar{M}_n(\underline{x})} \sum_{i=1}^n M_i^{(n)}(\underline{x}) h(x_{i,r+1}) h(x_{i,r+2}), & \bar{M}_n(\underline{x}) > 0 \\ 0 , & \text{otherwise} \end{cases}$$

is such that (5.4.19) holds. Therefore it is clear how to obtain an asymptotically optimal test of the hypothesis (2.3.13) with loss function (2.3.14).

5.5. Non-parametric Empirical Bayes Approach For Selecting the Best of k Populations.

The non-parametric empirical Bayes approach can be applied to the problem of selecting the best of k populations by dropping the previous assumptions of chapter four that f_{λ_i} and G_i are members of specific parametric families with unknown parameters. It will be shown that only mild assumptions on the first two moments of the conditional distribution of X_i given λ_i are necessary to derive empirical Bayes procedures which select the best of k populations when

(1) f_{λ} is discrete ; (2) f_{λ} is continuous.

Define the class \mathcal{G}_p to be

$$(5.5.1) \quad \mathcal{G}_p = \left\{ G = \prod_{i=1}^k G_i : \int_{\Omega} |\lambda_i|^p dG_i(\lambda_i) < \infty \text{ for all } i = 1, 2, \dots, k \right\} .$$

Suppose we have one observation per population; i.e. $r = 1$ in chapter four. Let G be in \mathcal{G}_p for some $p \geq 2$, and let the loss function be given by (4.2.2). It follows from the results of section 4.6 that if $\xi_{n,i}(x_i)$ is a function of the prior observations $x_{1,i}, x_{2,i}, \dots, x_{n,i}$ from the i 'th population such that

$$(5.5.2) \quad \xi_{n,i}(x_i) \xrightarrow{P} E(\Lambda_i | x_i)$$

for each $i = 1, 2, \dots, k$, then the empirical Bayes procedure for selecting the best of k populations with respect to any $G \in \mathcal{G}_p$ is given by

$$(5.5.3) \quad \delta_n(\underline{x}) = d_j \text{ where } j \text{ is any integer } 1, 2, \dots, k \text{ such that}$$

$$\xi_{n,j}(\underline{x}_j) = \max_{1 \leq i \leq k} \{\xi_{n,i}(\underline{x}_i)\}.$$

For each population, a function of the prior observations from that population which converges in probability to the a posteriori mean can be found when f_λ is discrete or continuous. Assume for both cases that

$$(5.5.4) \quad E(X_i | \Lambda_i) = \lambda_i,$$

and

$$(5.5.5) \quad E(X_i^2 | \Lambda_i) = c_1 + c_2 \lambda_i^p$$

for constants c_1, c_2 , and $p \geq 2$. The subscript "i" which indicated the i'th population will now be dropped for convenience since it is desired to find a consistent estimate for the a posteriori mean for a typical population.

Assume that each of the prior observations is a vector of two independent observations while λ remains a constant. Thus our prior independent observations are $\underline{X}_j = (X_{j,1}, X_{j,2})$ with unknown parameter λ_j for $j = 1, 2, \dots, n$. With appropriate notational changes, the

empirical Bayes procedure given by (5.5.3) is still valid. Using (5.5.1), (5.5.4), and (5.5.5) we have

$$\begin{aligned}
 (5.5.6) \quad E(X_{j,2} | X_{j,1}) &= E\{E(X_{j,2} | \Lambda_j, X_{j,1}) | X_{j,1}\} \\
 &= E\{E(X_{j,2} | \Lambda_j) | X_{j,1}\} \\
 &= E(\Lambda_j | X_{j,1})
 \end{aligned}$$

Since $X_{j,1}$ and $E(\Lambda_j | X_{j,1})$ have the same joint distribution as X and $E(\Lambda | X)$, then for any observation x of X ,

$$(5.5.7) \quad E(\Lambda_j | X_{j,1} = x) = E(\Lambda | X = x) .$$

Consider the cases where $f_\lambda(x)$ is discrete or continuous.

Discrete case. In this case $f_\lambda(x) > 0$ for every $x \in S$, a countable set, and for every $\lambda \in \Omega$ (5.5.7) suggests the following estimation procedure. Define for given $x \in S$ and each $j = 1, 2, \dots, n$,

$$(5.5.8) \quad Y_j = \frac{X_j}{n} = \begin{cases} X_{j,2} , & \text{if } X_{j,1} = x , \\ 0 , & \text{otherwise ,} \end{cases}$$

and

$$(5.5.9) \quad K_j = \begin{cases} 1 , & \text{if } X_{j,1} = x . \\ 0 , & \text{otherwise .} \end{cases}$$

Then

where \mathcal{H}_1 and \mathcal{H}_2 are Hilbert spaces and $\mathcal{H}_1 \otimes \mathcal{H}_2$ is the tensor product of \mathcal{H}_1 and \mathcal{H}_2 .

$$(\mathcal{H}_1 \otimes \mathcal{H}_2) \ni \psi \mapsto \psi \otimes \phi \in \mathcal{H}_1 \otimes \mathcal{H}_2 \quad (3.6.7)$$

$$(\mathcal{H}_1 \otimes \mathcal{H}_2) \ni \psi \mapsto \psi \otimes \phi \in \mathcal{H}_1 \otimes \mathcal{H}_2$$

$$(\mathcal{H}_1 \otimes \mathcal{H}_2) \ni \psi \mapsto \psi \otimes \phi \in \mathcal{H}_1 \otimes \mathcal{H}_2$$

where \mathcal{H}_1 and \mathcal{H}_2 are Hilbert spaces and $\mathcal{H}_1 \otimes \mathcal{H}_2$ is the tensor product of \mathcal{H}_1 and \mathcal{H}_2 .

$$(\mathcal{H}_1 \otimes \mathcal{H}_2) \ni \psi \mapsto \psi \otimes \phi \in \mathcal{H}_1 \otimes \mathcal{H}_2 \quad (3.6.8)$$

where \mathcal{H}_1 and \mathcal{H}_2 are Hilbert spaces and $\mathcal{H}_1 \otimes \mathcal{H}_2$ is the tensor product of \mathcal{H}_1 and \mathcal{H}_2 .

where \mathcal{H}_1 and \mathcal{H}_2 are Hilbert spaces and $\mathcal{H}_1 \otimes \mathcal{H}_2$ is the tensor product of \mathcal{H}_1 and \mathcal{H}_2 .

$$(\mathcal{H}_1 \otimes \mathcal{H}_2) \ni \psi \mapsto \psi \otimes \phi \in \mathcal{H}_1 \otimes \mathcal{H}_2 \quad (3.6.9)$$

where

$$(\mathcal{H}_1 \otimes \mathcal{H}_2) \ni \psi \mapsto \psi \otimes \phi \in \mathcal{H}_1 \otimes \mathcal{H}_2 \quad (3.6.10)$$

where

$$\begin{aligned}
 (5.5.10) \quad EY_j &= E(X_{j,2} | X_{j,1} = x) \cdot f_G(x) \\
 &= E(\Lambda_j | X_{j,1}) \cdot f_G(x) < \infty
 \end{aligned}$$

from (5.5.6) and where

$$f_G(x) = P(X = x) = \int_{\Omega} f_{\lambda}(x) \, dG(\lambda) \quad .$$

Also, using (5.5.5) and the observed value y of $X_{j,2}$, we have

$$\begin{aligned}
 EY_j^2 &\leq EX_{j,2}^2 \\
 &= \sum_S y^2 f_G(y) \\
 &= \sum_S y^2 \int_{\Omega} f_{\lambda}(y) \, dG(\lambda) \\
 &= \int_{\Omega} \left(\sum_S y^2 f_{\lambda}(y) \right) dG(\lambda) \\
 &\leq \int_{\Omega} (c_1 + c_2 \lambda^p) \, dG(\lambda) < \infty
 \end{aligned}$$

due to (5.5.1). Thus $\underline{X}_1, \underline{X}_2, \dots, \underline{X}_n$ are independent and identically distributed random variables with finite variance. By (5.5.7) and the strong law of large numbers,

$$(5.5.11) \quad \bar{Y} = \frac{1}{n} \sum_{j=1}^n Y_j \xrightarrow{a.s.} E(\Lambda | X = x) \cdot f_G(x) \quad .$$

Now

$$\sum_{j=1}^n Y_j > 0 \quad \text{implies that} \quad \sum_{j=1}^n K_j > 0$$

and we define

$$(5.5.12) \quad Z_n = \begin{cases} \frac{\sum_{j=1}^n Y_j}{\sum_{j=1}^n K_j} , & \text{if } \sum_{j=1}^n Y_j > 0 . \\ 0 & , \text{ otherwise } . \end{cases}$$

Since

$$(5.5.13) \quad \frac{1}{n} \sum_{j=1}^n K_j \xrightarrow{\text{a.s.}} f_G(x) > 0 ,$$

we have from (5.5.11) that

$$(5.5.14) \quad Z_n \xrightarrow{\text{a.s.}} E(\Lambda | X = x) .$$

Thus from (5.5.3) we see that the empirical Bayes procedure is given by

$$(5.5.15) \quad \delta_n(\underline{x}) = d_j \quad \text{where } j \text{ is any integer } 1, 2, \dots, k \text{ such}$$

$$\text{that } Z_{n,j} = \max_{1 \leq i \leq k} \{Z_{n,i}\} ,$$

where for each i , $Z_{n,i}$ is computed as Z_n in (5.5.12) for the

i'th population as the typical population.

Continuous case. Assume that $f_\lambda(x)$ is a continuous function on a set S for any $\lambda \in \Omega$ such that $f_\lambda(x) > 0$ for any $x \in S$ and

$$\int_S f_\lambda(x) d\mu(x) = 1 .$$

Assume also that there exists a real number K such that $f_\lambda(x) \leq K$ for all $x \in S$ and $\lambda \in \Omega$; i.e. f_λ is bounded uniformly. For each n partition the real axis into non-overlapping intervals $[\frac{q}{n^{1/2}}, \frac{q+1}{n^{1/2}})$, where $q = 0, \pm 1, \pm 2, \dots$. For the present observation $x \in S$, let $I^{(n)}(x)$ denote the unique subinterval which contains x for each $n = 1, 2, \dots$, and define

$$(5.5.16) \quad Y_{n,j} = \frac{X_{n,j}}{n^{1/2}} = \begin{cases} X_{j,2} , & \text{if } X_{j,1} \in I^{(n)}(x) \\ 0 , & \text{otherwise} \end{cases}$$

Now if

$$f_G(z, y) = P(X_{j,1} = y, X_{j,2} = z) = \int_\Omega f_\lambda(z) f_\lambda(y) dG(\lambda) ,$$

then

$$f_G(z) = \int_S f_G(z, y) d\mu(y) ,$$

and

$$(5.5.17) \quad \begin{aligned} EY_{n,j} &= \int_S n^{1/2} z \left(\int_{I^{(n)}(x)} f_G(z, y) d\mu(y) \right) d\mu(z) \\ &= b_n \end{aligned}$$

Then by Fubini's theorem,

$$(5.5.18) \quad \lim_{n \rightarrow \infty} b_n = \lim_{n \rightarrow \infty} \int_{\Omega} \left\{ \int_S z f_{\lambda}(z) d\mu(z) \right\} \left\{ n^{1/2} \int_{I^{(n)}(x)} f_{\lambda}(y) d\mu(y) \right\} dG(\lambda) ,$$

and from (5.5.4),

$$\int_S z f_{\lambda}(z) d\mu(z) = E(X_{j,2} | \Lambda) = \lambda .$$

Since f_{λ} is continuous, we have by the first mean value theorem and by the uniform boundedness of f_{λ} that there exists a $\xi_n \in I^{(n)}(x)$ such that

$$n^{1/2} \int_{I^{(n)}(x)} f_{\lambda}(y) d\mu(y) = f_{\lambda}(\xi_n) \leq K ,$$

and

$$\lim_{n \rightarrow \infty} n^{1/2} \int_{I^{(n)}(x)} f_{\lambda}(y) d\mu(y) = \lim_{n \rightarrow \infty} f_{\lambda}(\xi_n) = f_{\lambda}(x) .$$

Therefore from (5.5.18),

$$(5.5.19) \quad \lim_{n \rightarrow \infty} b_n = \int_{\Omega} \lambda f_{\lambda}(x) dG(\lambda) = E(\Lambda | X = x) : f_G(x) < \infty ,$$

since $f_{\lambda}(x) \leq K$ and $G \in \mathcal{G}_p$.

Define for each n and $j = 1, 2, \dots, n$ the random variable

$$(5.5.20) \quad Z_{n,j} = Y_{n,j} - E Y_{n,j} = Y_{n,j} - b_n$$

and the random sum

$$(5.5.21) \quad S_{n,n} = \sum_{j=1}^n Z_{n,j}.$$

Then

$$(5.5.22) \quad \begin{aligned} E Z_{n,j}^2 &= E Y_{n,j}^2 - 2b_n E Y_{n,j} + b_n^2 \\ &= E Y_{n,j}^2 - b_n^2, \end{aligned}$$

and

$$(5.5.23) \quad \begin{aligned} E Y_{n,j}^2 &\leq E X_{n,2}^2 \\ &= E \{E[X_{n,2}^2 | \Lambda_n]\} \\ &= E \{c_1 + c_2 \lambda_n^p\} \\ &= M < \infty \end{aligned}$$

for every $n = 1, 2, \dots$, since the $X_{j,n}$'s are independent and identically distributed for $j = 1, 2, \dots, n$, $G \in \mathcal{G}_p$, and by (5.5.5). From (5.5.19), (5.5.23), and the basic lemma on p. 277 of [9] which says that if

$$S_{n,n} = \sum_{j=1}^n Z_{n,j}$$

where the summands are independent random variables, and

$$\frac{1}{n^2} \sum_{j=1}^n E |Z_{n,j}|^2 \rightarrow 0, \quad \text{then} \quad \mathcal{L}\left(\frac{S_{n,n}}{n}\right) \rightarrow \mathcal{L}(0),$$

we have

$$(5.5.24) \quad \frac{S_{n,n}}{n} = \frac{1}{n} \sum_{j=1}^n Y_{n,j} \xrightarrow{P} 0.$$

This implies from (5.5.19) that

$$(5.5.25) \quad \bar{Y}_n = \frac{1}{n} \sum_{j=1}^n Y_{n,j} \xrightarrow{P} E(\Lambda | X = x) \cdot f_G(x).$$

Since $f_G(x) > 0$ for every $x \in S$, its consistent estimator $f_n(x)$ given by (4.6.10) with $r = 1$ is also strictly positive. Thus

$$(5.5.26) \quad \frac{\bar{Y}_n}{f_n(x)} \xrightarrow{P} E(\Lambda | X = x),$$

and from (5.5.3) the empirical Bayes procedure is given by

$$(5.5.27) \quad \delta_n(\underline{x}) = d_j \quad \text{where } j \text{ is any integer } 1, 2, \dots, k \text{ such}$$

$$\text{that } \frac{\bar{Y}_{n,j}}{f_{n,j}(\underline{x}_j)} = \max_{1 \leq i \leq k} \left\{ \frac{\bar{Y}_{n,j}}{f_{n,i}(\underline{x}_i)} \right\},$$

where for each i , $\bar{Y}_{n,i}$ is computed as \bar{Y}_n in (5.5.25) for the i 'th population as the typical population and $f_{n,i}(\underline{x}_i)$ is given by (4.6.10).

If the present observation x is also a vector with two

observations, then an average of the two estimators obtained by first suppressing one and then the other component of x could be used with the convergence in probability required for empirical Bayes estimators still being obtained. A generalization is possible to the case in which the present observation is a vector of dimension q , and the prior observations have dimension $q+1$.

CHAPTER VI

COMPOUND DECISION PROBLEM

The compound decision problem, which is closely related to the empirical Bayes problem, occurs when one is confronted with n individual decisions about some unknown parameters $\lambda_1, \lambda_2, \dots, \lambda_n$. The parameters are considered to be an unknown sequence of constants and not as the realizations of n independent random variables Λ_i with distribution function G . Information concerning the frequency distribution of the parameters is obtained from the sequence of observations x_1, x_2, \dots , and the aim here is to approximate to the use of the decision function which would be optimal if the frequency distribution of the parameters were known in advance.

Early work on this problem dealt with the case in which the component decisions are of the simple versus simple hypothesis testing type and can therefore be stated in terms of testing whether $\lambda = 0$ or $\lambda = 1$. If decision functions are allowed to depend on the data from all n components, we have the nonsequential case which was considered by Hannan and Robbins [4]. Samuel [19], [21] considers the sequential case where it is assumed that at the time a decision is made in any particular component problem the available information includes the data obtained in all previous component decision problems in the sequence.

Let X be a random variable which is known to have one of two distinct distributions $F_\lambda(x)$ for $\lambda = 0$ or 1 . On the basis of a

single observation of X it is required to decide whether the true value of the unknown parameter is 0 or 1 .

Consider now statistical decision problems of the same formal structure which are in a large group. That is, we have a sequence of independent random variables X_1, X_2, \dots and parameter values $\lambda_1, \lambda_2, \dots$ where each λ_i is 0 or 1 , and X_i has the distribution function $F_{\lambda_i}(x_i)$ which is given in terms of known densities $f_{\lambda_i}(x_i)$ with respect to some measure μ . The sequence of λ 's is unknown, and we are required to decide, for each $i = 1, 2, \dots$ whether $\lambda_i = 0$ or 1 .

In the sequential case the decision about λ_i may depend on the observed values $\underline{x}_i = (x_1, x_2, \dots, x_i)$ of $\underline{X}_i = (X_1, X_2, \dots, X_i)$. The possible actions for any decision are d_1 , where the experimenter says " $\lambda = 0$ ", and d_2 , where the experimenter says " $\lambda = 1$ " . Our loss structure is defined by

$$(6.1) \quad \begin{array}{ll} L(d_1, \lambda = 0) = 0 & L(d_1, \lambda = 1) = a \\ L(d_2, \lambda = 0) = b & L(d_2, \lambda = 1) = 0 \end{array} ,$$

where $a > 0$ and $b > 0$.

Let δ be a decision function for the component problem. Then δ is a measurable function with $0 \leq \delta(x) \leq 1$, where, when $X = x$ is observed, one says " $\lambda = 1$ " with probability $\delta(x)$ and " $\lambda = 0$ " with probability $1 - \delta(x)$. Then the risk of δ is given by

$$(6.2) \quad R(\delta, \lambda) = \begin{cases} b \ E_{\lambda}(\delta(X)) , & \text{for } \lambda = 0 , \\ a \ E_{\lambda}(1 - \delta(X)) , & \text{for } \lambda = 1 , \end{cases}$$

where E_{λ} denotes expectation with respect to F_{λ} for $\lambda = 0, 1$.

If λ is the realization of a random variable Λ with the "a priori distribution" $P(\Lambda = 1) = \eta = 1 - P(\Lambda = 0)$, then the overall expected loss will be

$$\begin{aligned} (6.3) \quad R(\delta, \eta) &= \eta R(\delta, 1) + (1-\eta) R(\delta, 0) \\ &= \eta a \int (1 - \delta(x)) f_1(x) d\mu(x) \\ &\quad + (1-\eta) b \int \delta(x) f_0(x) d\mu(x) \\ &= \int [(1-\eta) b f_0(x) - \eta a f_1(x)] \delta(x) d\mu(x) \\ &\quad + \int \eta a f_1(x) d\mu(x) \\ &= \eta a + \int [(1-\eta) b f_0(x) - \eta a f_1(x)] \delta(x) d\mu(x) . \end{aligned}$$

For fixed η (6.3) is minimized with respect to all possible decision functions δ by any δ_{η} of the form

$$(6.4) \quad \delta_{\eta} = \begin{cases} 1 & , \text{ if } (1-\eta) b f_0(x) < \eta a f_1(x) . \\ 0 & , \text{ if } (1-\eta) b f_0(x) > \eta a f_1(x) . \\ \text{arbitrary in } [0, 1] , & \text{ if } (1-\eta) b f_0(x) = \eta a f_1(x) . \end{cases}$$

Denote by δ_{η}^0 the non-randomized function

$$(6.5) \quad \delta_{\eta}^0 = \begin{cases} 1, & \text{if } (1-\eta)bf_0(x) < \eta af_1(x) \\ 0, & \text{otherwise.} \end{cases}$$

The rules δ_{η} which minimize (6.3) are called Bayes with respect to the a priori distribution η , and the function

$$(6.6) \quad R(\eta) = R(\delta_{\eta}, \eta) = \min_{\delta} R(\delta, \eta)$$

is called Bayes envelope function.

Returning to our sequence of decision problems, let

$\Omega^{\infty} = \{\text{all possible infinite sequences of 0's and 1's}\}$, and

$\Omega^n = \{\text{all } 2^n \text{ n-vectors of 0's and 1's}\}$. For any

$\underline{\lambda} = (\lambda_1, \lambda_2, \dots) \in \Omega^{\infty}$, let $\underline{\lambda}_n = (\lambda_1, \lambda_2, \dots, \lambda_n) \in \Omega^n$ denote its initial n-vector with corresponding random variable $\underline{X}_n = (X_1, X_2, \dots, X_n)$ where X_i is distributed according to F_{λ_i} with density $f_{\lambda_i}(x_i)$ and is independent of the other X 's and λ 's. No relationship among the λ 's is assumed.

By an (n-step) compound decision rule we mean any n-vector

$D_n = (\delta_1, \delta_2, \dots, \delta_n)$ of measurable functions where, in the sequential case, $\delta_i = \delta_i(x_1, x_2, \dots, x_i)$ and where $0 \leq \delta_i \leq 1$ is the probability with which one decides $\lambda_i = 1$ when $\underline{X}_i = \underline{x}_i$ has been observed. The risk of D_n at the point $\underline{\lambda}_n$ is defined by

$$(6.7) \quad R(D_n, \underline{\lambda}_n) = \frac{1}{n} \sum_{i=1}^n R(\delta_i, \underline{\lambda}_i),$$

Let $f(x) = \frac{1}{x}$ be the function.

$$f(x) = \frac{1}{x} \Rightarrow f'(x) = -\frac{1}{x^2} \Rightarrow f'(1) = -1$$

The value of $f'(1)$ is -1 . The value of $f(1)$ is 1 . The value of $f(2)$ is $\frac{1}{2}$. The value of $f(3)$ is $\frac{1}{3}$. The value of $f(4)$ is $\frac{1}{4}$. The value of $f(5)$ is $\frac{1}{5}$. The value of $f(6)$ is $\frac{1}{6}$. The value of $f(7)$ is $\frac{1}{7}$. The value of $f(8)$ is $\frac{1}{8}$. The value of $f(9)$ is $\frac{1}{9}$. The value of $f(10)$ is $\frac{1}{10}$.

$$f(x) = \frac{1}{x} \Rightarrow f'(x) = -\frac{1}{x^2} \Rightarrow f'(1) = -1$$

is the value of $f'(1)$.

Let $f(x) = \frac{1}{x}$ be the function.

$$f(x) = \frac{1}{x} \Rightarrow f'(x) = -\frac{1}{x^2} \Rightarrow f'(1) = -1$$

$$f(x) = \frac{1}{x} \Rightarrow f'(x) = -\frac{1}{x^2} \Rightarrow f'(1) = -1$$

$$f(x) = \frac{1}{x} \Rightarrow f'(x) = -\frac{1}{x^2} \Rightarrow f'(1) = -1$$

$$f(x) = \frac{1}{x} \Rightarrow f'(x) = -\frac{1}{x^2} \Rightarrow f'(1) = -1$$

$$f(x) = \frac{1}{x} \Rightarrow f'(x) = -\frac{1}{x^2} \Rightarrow f'(1) = -1$$

$$f(x) = \frac{1}{x} \Rightarrow f'(x) = -\frac{1}{x^2} \Rightarrow f'(1) = -1$$

is the value of $f'(1)$.

Let $f(x) = \frac{1}{x}$ be the function.

$$f(x) = \frac{1}{x} \Rightarrow f'(x) = -\frac{1}{x^2} \Rightarrow f'(1) = -1$$

$$f(x) = \frac{1}{x} \Rightarrow f'(x) = -\frac{1}{x^2} \Rightarrow f'(1) = -1$$

$$f(x) = \frac{1}{x} \Rightarrow f'(x) = -\frac{1}{x^2} \Rightarrow f'(1) = -1$$

$$f(x) = \frac{1}{x} \Rightarrow f'(x) = -\frac{1}{x^2} \Rightarrow f'(1) = -1$$

is the value of $f'(1)$.

$$f(x) = \frac{1}{x} \Rightarrow f'(x) = -\frac{1}{x^2} \Rightarrow f'(1) = -1$$

is

where

$$(6.8) \quad R(\delta_i, \lambda_i) = \int [a\lambda_i(1-\delta_i(\underline{x}_i)) + b(1-\lambda_i)\delta_i(\underline{x}_i)] f_{\lambda_i}(\underline{x}_i) d\mu^i.$$

Denote the random loss incurred in the i 'th decision by $L(\delta_i(\underline{x}_i), \lambda_i)$. Corresponding to (6.7) we have

$$(6.9) \quad L(D_n, \lambda_n) = \frac{1}{n} \sum_{i=1}^n L(\delta_i(\underline{x}_i), \lambda_i).$$

Let $h(x)$ be an unbiased estimate of λ , and define for $i = 1, 2, \dots$,

$$(6.10) \quad p_i = p_i(\underline{x}_i) = \begin{cases} 0 & , \text{ if } \frac{1}{i} \sum_{j=1}^i h(x_j) \leq 0, \\ \frac{1}{i} \sum_{j=1}^i h(x_j) & , \text{ if } 0 \leq \frac{1}{i} \sum_{j=1}^i h(x_j) \leq 1, \\ 1 & , \text{ if } 1 \leq \frac{1}{i} \sum_{j=1}^i h(x_j), \end{cases}$$

and set $p_0 = \frac{1}{2}$. Consider the rule with

$$(6.11) \quad \delta_i^*(\underline{x}_i) = \delta_{p_{i-1}}^0(\underline{x}_i) = \begin{cases} 1, & \text{ if } f_0(\underline{x}_i)b(1-p_{i-1}(\underline{x}_{i-1})) < f_1(\underline{x}_i)ap_{i-1}(\underline{x}_{i-1}), \\ 0, & \text{ otherwise,} \end{cases}$$

and let $D_n^* = (\delta_1^*, \delta_2^*, \dots, \delta_n^*)$. Letting

$$(6.12) \quad v_i = \frac{1}{i} \sum_{j=1}^i \lambda_j,$$

where

$$C_1 = \frac{1}{2} \left(\frac{1}{\alpha} + \frac{1}{\beta} \right) \quad (2.1)$$

and $C_2 = \frac{1}{2} \left(\frac{1}{\alpha} - \frac{1}{\beta} \right)$ is a constant.

$$C_3 = \frac{1}{2} \left(\frac{1}{\alpha} + \frac{1}{\beta} \right) \quad (2.2)$$

$$C_4 = \frac{1}{2} \left(\frac{1}{\alpha} - \frac{1}{\beta} \right) \quad (2.3)$$

and $C_5 = \frac{1}{2} \left(\frac{1}{\alpha} + \frac{1}{\beta} \right)$ is a constant.

$$C_6 = \frac{1}{2} \left(\frac{1}{\alpha} - \frac{1}{\beta} \right)$$

$$C_7 = \frac{1}{2} \left(\frac{1}{\alpha} + \frac{1}{\beta} \right)$$

$$\left. \begin{aligned} C_8 &= \frac{1}{2} \left(\frac{1}{\alpha} - \frac{1}{\beta} \right) \\ C_9 &= \frac{1}{2} \left(\frac{1}{\alpha} + \frac{1}{\beta} \right) \\ C_{10} &= \frac{1}{2} \left(\frac{1}{\alpha} - \frac{1}{\beta} \right) \end{aligned} \right\} \quad (2.4)$$

where $C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8, C_9, C_{10}$ are constants.

$$C_{11} = \frac{1}{2} \left(\frac{1}{\alpha} + \frac{1}{\beta} \right) \quad (2.5)$$

$$C_{12} = \frac{1}{2} \left(\frac{1}{\alpha} - \frac{1}{\beta} \right)$$

$$C_{13} = \frac{1}{2} \left(\frac{1}{\alpha} + \frac{1}{\beta} \right) \quad (2.6)$$

D_n^* uses a rule which is Bayes with respect to the estimate p_{i-1} of v_{i-1} to decide on λ_i . Samuel proves (Theorem 1 in [19]) that if $R(\eta)$ has a derivative for $0 \leq \eta \leq 1$, then $\{D_n^*\}$ defined by (6.11) is such that for every $\epsilon > 0$ there exists an $N(\epsilon)$ such that for all $n \geq N(\epsilon)$,

$$(6.13) \quad R(D_n^*, \lambda_n) - R(v_n) < \epsilon \quad \text{uniformly in } \lambda_n \in \Omega^n.$$

Thus even if the sequence $\lambda_1, \lambda_2, \dots$ is chosen arbitrarily by Nature, our average performance in using D^* on the first n decisions will, for large n , be almost as good as if we had used throughout the Bayes decision function δ_v corresponding to the actual proportion

$$v = \frac{1}{n} \sum_{i=1}^n \lambda_i$$

of ones among $\lambda_1, \lambda_2, \dots, \lambda_n$.

Suppose that $R'(\eta)$ does not exist for some value η . By means of randomization using a uniformly distributed random variable, Samuel finds decision rules whose Bayes envelope function is some "smoothed" version of the original $R(\eta)$ and stays sufficiently close to $R(\eta)$. Let $\underline{Z} = (z_1, z_2)$ be a random variable which is uniformly distributed over the unit square. Assuming that all \underline{Z} 's are independent of the other random variables X_i , $i = 1, 2, \dots, n$, let

$$(6.14) \quad m(\underline{Z}, \underline{x}_i) = \frac{p_i(\underline{x}_i) + i^{-1/4} z_2}{1 + i^{-1/4} (z_1 + z_2)}, \quad i = 1, 2, \dots,$$

where $p_i(\underline{x}_i)$ is defined in (6.10), and $m(\underline{Z}, \underline{x}_0) = \frac{1}{2}$. Consider the rule with

$$(6.15) \quad \hat{\delta}_i(\underline{x}_i) = \delta_{m_{i-1}}^0(\underline{x}_i),$$

where $m_{i-1} = m(\underline{Z}, \underline{x}_{i-1})$, and let $\hat{D}_n = (\hat{\delta}_1(\underline{x}_1), \hat{\delta}_2(\underline{x}_2), \dots, \hat{\delta}_n(\underline{x}_n))$ for $n = 1, 2, \dots$. Samuel proves (Theorem 2 in [19]) that the sequence of rules $\{\hat{D}_n\}$ is such that for every $\epsilon > 0$ there exists an $N(\epsilon)$ such that for all $n \geq N(\epsilon)$,

$$(6.16) \quad R(\hat{D}_n, \underline{\lambda}_n) - R(v_n) < \epsilon \quad \text{uniformly in } \underline{\lambda}_n \in \Omega^n.$$

\hat{D}_n is not an admissible rule since no admissible rule involves artificial randomization of the above kind. \hat{D}_n is not optimal in any sense other than being optimal in the limit.

Samuel [21] proves analogous results to (6.13) and (6.16) when the loss given by (6.9) is used. It is shown (Theorems 1 and 2 in [21]) that for any sequence of λ -values, the difference between the loss incurred by \hat{D}_n , $L(\hat{D}_n, \underline{\lambda}_n)$, and $R(v_n)$ converges to zero in probability, and if $R(\eta)$ has a derivative for $0 \leq \eta \leq 1$, a corresponding statement holds with probability one for D_n^* .

For the nonsequential case where all n random variables X_1, X_2, \dots, X_n may be observed before the decisions on λ_i , $i = 1, 2, \dots, n$ have to be made, Hannan and Robbins derive (Theorem 3 in [4]) a decision rule for which $P_{\underline{\lambda}}(L(\underline{\lambda}_n) - R(v_n) \leq \epsilon \text{ for all } n > N) > 1 - \epsilon$ uniformly

in $\underline{\lambda} \in \Omega^\infty$, where $L(\underline{\lambda}_n)$ denotes the random loss incurred by their rule in the first n decisions. Hannan and Robbins also show (Theorem 4 in [4]) that for the nonsequential case there exists a sequence of rules $\{D_n\}$ with the property that given any $\epsilon > 0$ there exists an $N(\epsilon)$ such that for every $n > N(\epsilon)$,

$$(6.17) \quad R(D_n, \underline{\lambda}_n) - R(v_n) < \epsilon \text{ uniformly in } \underline{\lambda}_n \in \Omega^n,$$

which corresponds to (6.16) in the sequential case.

Van Ryzin [24] generalizes and strengthens the result (6.16) in the nonsequential problem to the case where each component problem consists of making one of n decisions based on an observation from one of m distributions. Let X be a random variable which is known to have one of m possible distributions F_λ where λ is in the finite parameter space $\Omega = \{1, 2, \dots, m\}$. After observing X a decision $d \in \mathcal{D}$ is made with loss $L(\lambda, d)$ if decision d is made when X is distributed as F_λ where $\lambda = 1, 2, \dots, m$, and $d = 1, 2, \dots, n$. Thus we have an $m \times n$ loss matrix $L(\lambda, d)$. If X_i , $i = 1, 2, \dots, N$ are N independent observations with X_i distributed according to F_{λ_i} with $\lambda_i \in \Omega$, then a decision $d_i \in \mathcal{D}$ based on all N observations is to be made for each of the N component problems. This is a finite compound decision problem.

Let there exist a σ -finite measure μ which dominates $\{F_1, F_2, \dots, F_m\}$ such that the densities

$$(6.18) \quad f_{\lambda}(x) = \left(\frac{dF_{\lambda}}{d\mu} \right)(x) \leq K \quad \text{a.e. } (\mu)x$$

for some $K < \infty$, and let $f(x) = (f_1(x), f_2(x), \dots, f_m(x))$ be the vector of densities in (6.18).

A randomized decision function for the compound decision problem will be any $N \times n$ matrix of measurable functions $D(\underline{x}) = (\delta_d^k(\underline{x}))$ such that for $k = 1, 2, \dots, N$ and $d = 1, 2, \dots, n$,

$$\delta_d^k(\underline{x}) = P\{d_k = d | X = \underline{x}\}, \quad \text{and} \quad \sum_{d=1}^n \delta_d^k = 1,$$

where $\underline{x} = (x_1, x_2, \dots, x_N)$ is the vector of observations. Denote the k 'th row of $D(\underline{x})$ by $\delta^{(k)}(\underline{x}) = (\delta_1^k(\underline{x}), \delta_2^k(\underline{x}), \dots, \delta_n^k(\underline{x}))$. Let \mathfrak{X} be the set of all N -tuples $\underline{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_N)$ where $\lambda_k \in \Omega$. For the $m \times n$ matrix of losses $L(\lambda, d)$, the rows will be denoted by L_{λ} and the columns by L^d for $\lambda = 1, 2, \dots, m$, and $d = 1, 2, \dots, n$. If E^m is m -dimensional Euclidean space, then if $y = (y_1, y_2, \dots, y_m)$ and $z = (z_1, z_2, \dots, z_m)$ are vectors in E^m , then the vector $yz = (y_1 z_1, y_2 z_2, \dots, y_m z_m)$, and the inner product is

$$(y, z) = \sum_{i=1}^m y_i z_i.$$

The risk function $R(\underline{\lambda}, D)$ for the compound decision procedure $D(\underline{x})$ is defined to be the average of the component risks

$$f(x) = \frac{1}{2} \left(x + \frac{1}{x} \right) \quad (1.1)$$

Let us consider the function $f(x)$ defined on the interval $(0, \infty)$ by the formula (1.1). It is easy to see that $f(x) > 1$ for all $x > 0$.

Let us also consider the function $g(x)$ defined on the interval $(0, \infty)$ by the formula

$$g(x) = \frac{1}{2} \left(x - \frac{1}{x} \right) \quad (1.2)$$

It is easy to see that $g(x) < 1$ for all $x > 0$.

$$f(x) = \frac{1}{2} \left(x + \frac{1}{x} \right) \quad (1.3)$$

Let us consider the function $h(x)$ defined on the interval $(0, \infty)$ by the formula (1.3). It is easy to see that $h(x) > 1$ for all $x > 0$. Let us also consider the function $k(x)$ defined on the interval $(0, \infty)$ by the formula (1.4). It is easy to see that $k(x) < 1$ for all $x > 0$.

Let us consider the function $l(x)$ defined on the interval $(0, \infty)$ by the formula (1.5). It is easy to see that $l(x) > 1$ for all $x > 0$.

Let us consider the function $m(x)$ defined on the interval $(0, \infty)$ by the formula (1.6). It is easy to see that $m(x) < 1$ for all $x > 0$.

Let us consider the function $n(x)$ defined on the interval $(0, \infty)$ by the formula (1.7). It is easy to see that $n(x) > 1$ for all $x > 0$.

$$f(x) = \frac{1}{2} \left(x + \frac{1}{x} \right) \quad (1.8)$$

Let us consider the function $o(x)$ defined on the interval $(0, \infty)$ by the formula (1.9). It is easy to see that $o(x) > 1$ for all $x > 0$.

Let us consider the function $p(x)$ defined on the interval $(0, \infty)$ by the formula (1.10). It is easy to see that $p(x) < 1$ for all $x > 0$.

$$(6.19) \quad R_k(\underline{\lambda}, D) = E \left\{ \sum_{d=1}^n L(\lambda_k, d) \delta_d^k(\underline{x}) \right\} \\ = E(L_{\lambda_k}, \delta^{(k)}(\underline{x})) ,$$

for each subproblem $k = 1, 2, \dots, N$ where E denotes expectation with respect to $\prod_{k=1}^N F_{\lambda_k}$. Thus

$$(6.20) \quad R(\underline{\lambda}, D) = \frac{1}{N} \sum_{k=1}^N R_k(\underline{\lambda}, D) = E \left[\frac{1}{N} \sum_{k=1}^N (L_{\lambda_k}, \delta^{(k)}(\underline{x})) \right] .$$

The empirical distribution on Ω is given by the vector $p(\underline{\lambda}) = (p_1(\underline{\lambda}), p_2(\underline{\lambda}), \dots, p_m(\underline{\lambda}))$, where for $\underline{\lambda} \in \mathfrak{N}$ and $\lambda = 1, 2, \dots, m$ we define

$$(6.21) \quad p_{\lambda}(\underline{\lambda}) = \frac{1}{N} (\text{number of } \lambda_k = \lambda, \quad k \leq N) .$$

Suppose that the decision function $D(\underline{x})$ is simple; i.e. there exist functions $\delta_d(\cdot)$, $d = 1, 2, \dots, n$ such that $\delta^{(k)}(\underline{x}) = (\delta_1(x_k), \delta_2(x_k), \dots, \delta_n(x_k))$ for $k = 1, 2, \dots, N$, and we write $\delta = (\delta_1, \delta_2, \dots, \delta_n)$. Then $\delta_d(x) \geq 0$ for $d = 1, 2, \dots, n$ is a set of measurable functions such that

$$\sum_{d=1}^n \delta_d(x) = 1 .$$

The risk incurred in using procedure δ is from (6.20)

$$\int_{\mathbb{R}^n} |\nabla u|^2 dx = \int_{\mathbb{R}^n} |\nabla v|^2 dx \quad (1.1)$$

$$(\nabla u, \nabla v) = 0$$

Let us assume that $\nabla u \cdot \nabla v = 0$ almost everywhere in \mathbb{R}^n .

$$\int_{\mathbb{R}^n} |\nabla u|^2 dx = \int_{\mathbb{R}^n} |\nabla v|^2 dx$$

$$(\nabla u, \nabla v) = \int_{\mathbb{R}^n} \nabla u \cdot \nabla v dx = 0 \quad (1.2)$$

The equality (1.2) shows that ∇u and ∇v are orthogonal in the sense of the inner product in $L^2(\mathbb{R}^n)$. This implies that ∇u and ∇v are linearly independent.

$$\int_{\mathbb{R}^n} |\nabla u|^2 dx = \int_{\mathbb{R}^n} |\nabla v|^2 dx \quad (1.3)$$

Since ∇u and ∇v are linearly independent, we have $\nabla u \neq 0$ and $\nabla v \neq 0$ almost everywhere in \mathbb{R}^n .

$$\int_{\mathbb{R}^n} |\nabla u|^2 dx = \int_{\mathbb{R}^n} |\nabla v|^2 dx \quad (1.4)$$

Let us assume that $\nabla u \cdot \nabla v = 0$ almost everywhere in \mathbb{R}^n .

$$\int_{\mathbb{R}^n} |\nabla u|^2 dx = \int_{\mathbb{R}^n} |\nabla v|^2 dx \quad (1.5)$$

$$\int_{\mathbb{R}^n} |\nabla u|^2 dx = \int_{\mathbb{R}^n} |\nabla v|^2 dx$$

The equality (1.5) shows that ∇u and ∇v are linearly independent.

$$\begin{aligned}
 R(\underline{\lambda}, \delta) &= \frac{1}{N} \sum_{k=1}^N E(L_{\lambda_k}, \delta(X_k)) \\
 &= \sum_{\lambda=1}^m p_{\lambda}(\underline{\lambda}) E_{\lambda}(L_{\lambda}, \delta(X)) \\
 (6.22) \quad &= \sum_{\lambda=1}^m p_{\lambda}(\underline{\lambda}) E_{\lambda} \left\{ \sum_{d=1}^n L(\lambda, d) \delta_d(X) \right\} \\
 &= \sum_{\lambda=1}^m p_{\lambda}(\underline{\lambda}) \rho_{\lambda}(\delta) \\
 &= (p(\underline{\lambda}), \rho(\delta)) \quad ,
 \end{aligned}$$

where

$$\rho(\delta) = E_{\lambda}(L_{\lambda}, \delta(X)) = E_{\lambda} \left\{ \sum_{d=1}^n L(\lambda, d) \delta_d(X) \right\} \quad ,$$

E_{λ} denoting expectation with respect to F_{λ} , and $\rho(\delta) = (\rho_1(\delta), \rho_2(\delta), \dots, \rho_n(\delta))$.

If $\xi = (\xi_1, \xi_2, \dots, \xi_m)$ is any vector in E^m , let

$$(6.23) \quad \psi(\xi, \delta) = (\xi, \rho(\delta)) \quad .$$

Thus when $\xi = p(\underline{\lambda})$, $\psi(\xi, \delta)$ is the risk function for δ given in (6.22). From (6.18) and (6.23) we have

$$(6.24) \quad \psi(\xi, \delta) = \sum_{\lambda=1}^m \xi_{\lambda} E_{\lambda} \left\{ \sum_{d=1}^n L(\lambda, d) \delta_d(X) \right\}$$

$$\begin{aligned}
 &= \int \sum_{\lambda=1}^m \sum_{d=1}^n \xi_{\lambda} L(\lambda, d) f_{\lambda}(x) \delta_d(x) d\mu(x) \\
 &= \int \left\{ \sum_{d=1}^n (\xi, L^d f(x)) \delta_d(x) \right\} d\mu(x) .
 \end{aligned}$$

For fixed ξ , (6.24) will be a minimum for any vector function δ_{ξ} , of the form

$$(6.25) \quad \delta_{\xi, d}(x) = \begin{cases} 0 & , \text{ if } (\xi, L^d f(x)) > \min_{1 \leq j \leq n} (\xi, L^j f(x)) \\ 1 & , \text{ if } (\xi, L^d f(x)) < \min_{j \neq d} (\xi, L^j f(x)) \\ \text{arbitrary,} & \text{if } (\xi, L^d f(x)) = \min_{j \neq d} (\xi, L^j f(x)) \end{cases}$$

such that $\delta_{\xi, d}(x) \geq 0$ for $d = 1, 2, \dots, n$, and

$$\sum_{d=1}^n \delta_{\xi, d}(x) = 1 \quad \text{a.e. } (\mu)x .$$

Thus if ξ is a bona fide a priori distribution, then such a δ_{ξ} would be a decision procedure Bayes against ξ .

Any randomized procedure of the form (6.25) may be replaced by the non-randomized version

$$(6.26) \quad \delta'_{\xi, d}(x) = \begin{cases} 1, & \text{if } d \text{ is the smallest integer for which} \\ & (\xi, L^d f(x)) = \min_{1 \leq j \leq n} (\xi, L^j f(x)), \\ 0, & \text{otherwise,} \end{cases}$$

which also minimizes $\psi(\xi, \delta)$ for fixed ξ .

If $L_1(\mu)$ and $L_2(\mu)$ are the function spaces of μ -integrable and μ -square integrable functions respectively, then $f_\lambda(x)$ for $\lambda = 1, 2, \dots, m$ are in $L_1(\mu)$ and $L_2(\mu)$ since they are bounded. Define

$$S^{(m)} = \left\{ \eta \mid \eta \in E^m, \eta_\lambda > 0, \sum_{\lambda=1}^m \eta_\lambda = 1 \right\}$$

to be the simplex in E^m , and for $\eta \in S^{(m)}$ define the probability mixture

$$F_\eta = \sum_{\lambda=1}^m \eta_\lambda F_\lambda$$

with μ -density $f_\eta(x) = (\eta, f(x))$, and let $\mathcal{F} = \{F_\eta \mid \eta \in S^{(m)}\}$ be the class of all mixtures.

A vector function $h = (h_1(x), h_2(x), \dots, h_m(x))$ into E^m with coordinate functions $h_j \in L_1(\mu)$ is an unbiased estimate for the class \mathcal{F} if $E_\eta\{h(X)\} = \eta$ for all $\eta \in S^{(m)}$, where E_η denotes expectation with respect to the mixture F_η , and if h exists, the class \mathcal{F} is called estimable. Denote the class of all unbiased estimates for the class \mathcal{F} by \mathcal{E} and the subclass of \mathcal{E} for which $h_j \in L_2(\mu)$ for $j = 1, 2, \dots, m$, by \mathcal{K} .

Define the random variable

$$(6.27) \quad \bar{h}(\underline{X}) = \frac{1}{N} \sum_{k=1}^N h(X_k),$$

where $h \in \mathcal{E}$ and $\underline{X} = (X_1, X_2, \dots, X_n)$. Van Ryzin shows that $\bar{h}(\underline{X})$ is an unbiased estimate of the empirical distribution $p(\underline{\lambda})$ for all $\underline{\lambda} \in \tilde{\Omega}$ and for $h \in \mathcal{K}$ defines the non-simple, non-randomized decision function

$$(6.28) \quad \delta'_{\bar{h},d}(\underline{x}_k) = \begin{cases} 1, & \text{if } d \text{ is the smallest integer for which} \\ & (\bar{h}, L^d f(\underline{x}_k)) = \min_{1 \leq j \leq n} (\bar{h}, L^j f(\underline{x}_k)) \\ 0, & \text{otherwise} \end{cases}$$

for $d = 1, 2, \dots, n$ which results from substituting $\bar{h}(\underline{X})$ for $p(\underline{\lambda})$ in (6.26). The resulting non-simple, non-randomized decision procedure D' consists of the N vector functions

$$\delta^k(\underline{x}) = \delta'_{\bar{h}}(\underline{x}_k) = (\delta'_{\bar{h},1}(\underline{x}_k), \delta'_{\bar{h},2}(\underline{x}_k), \dots, \delta'_{\bar{h},n}(\underline{x}_k))$$

for $k = 1, 2, \dots, N$. Let $\phi(p(\underline{\lambda})) = \inf_{\delta} \psi(p(\underline{\lambda}), \delta) = \inf_{\delta} R(\underline{\lambda}, \delta)$. Then

Van Ryzin proves (Theorem 2 in [24]) that if $h \in \mathcal{E}$ and

$E_{\lambda} |h_j(X)|^3 < \infty$ for λ , $j = 1, 2, \dots, m$, then

$$(6.29) \quad R(\underline{\lambda}, D') - \phi(p(\underline{\lambda})) \leq c N^{-1/2},$$

where c is independent of $\underline{\lambda} \in \tilde{\Omega}$ for all N . Thus Van Ryzin has found sufficient conditions for a uniform bound on the difference in risks (the regret function) of a certain compound procedure and a best "simple" procedure which is Bayes against the empirical distribution on the component parameter space.

In the above results the family of distribution functions governing the observation is assumed to be finite, and the main results are concerned only with the convergence to zero of the difference between the average risk and a certain "optimal" goal as the number of component problems becomes large.

In more recent work Swain [23] considers standard (infinite state) estimation problems with squared error loss. Let $\underline{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_n, \dots)$ be a countably infinite vector whose components λ_i are elements of some finite interval Ω of the real line; i.e. $-\infty < \alpha \leq \lambda_i \leq \beta < \infty$ for $i = 1, 2, \dots$. Let $\mathcal{F} = \{f_\lambda(x) : \lambda \in \Omega\}$ be a family of known probability density functions with parameter λ . If $\underline{\lambda}$ is unknown, we want to estimate λ_i for each i with each estimate being based on the independent observations X_j , $j = 1, 2, \dots, i$. Thus for each $i = 1, 2, \dots$, a non-randomized estimator $\phi_i(\underline{x}_i)$ is sought for λ_i where $\underline{x}_i = (x_1, x_2, \dots, x_i)$ is the vector of observations.

Assuming that we have a squared error loss, the loss when ϕ_i is an estimate of λ_i is $(\phi_i - \lambda_i)^2$. The risk of the estimator ϕ_i is defined to be the expected loss $E(\phi_i(\underline{X}_i) - \lambda_i)^2$, and the average risk for n estimations is then

$$\frac{1}{n} \sum_{i=1}^n E(\phi_i(\underline{X}_i) - \lambda_i)^2.$$

For specified Ω and \mathcal{F} a decision procedure $\phi = (\phi_1, \phi_2, \dots)$ is found by Swain which, on the basis of its average risk for the first n estimations, is in some sense optimal for large n . Samuel [22] also

deals with the sequential compound decision problem when the component problem is an estimation problem.

BIBLIOGRAPHY

1. Cramér, H. (1946). Mathematical Methods of Statistics, Princeton University Press.
2. Deely, J. J. (1965). "Multiple Decision Procedures from an Empirical Bayes Approach." Statistics Department, Purdue University Technical Report.
3. _____. (1965). "Non-parametric Empirical Bayes Procedures for Selecting the Best of k Populations." To appear in Annals of Mathematical Statistics.
4. Hannan, J. F. and Robbins, H. (1955). "Asymptotic Solutions of the Compound Decision Problem for Two Completely Specified Distributions." Annals of Mathematical Statistics, 26, p. 37-51.
5. Hannan, J. F. and Van Ryzin, J. R. (1965). "Rate of Convergence in the Compound Decision Problem for Two Completely Specified Distributions." Annals of Mathematical Statistics, 36, p. 1743-1752.
6. Johns, Jr., M. V. (1957). "Non-parametric Empirical Bayes Procedures." Annals of Mathematical Statistics, 28, p. 649-669.
7. _____. (1966). "Two-Action Compound Decision Problems." Stanford University Technical Report No. 87.
8. Krutchkoff, R. G. (1965). "A Supplementary Sample Non-parametric Empirical Bayes Approach to Some Problems in Statistical Decision Theory." Unpublished.
9. Loève, M. (1960). Probability Theory, second edition, Van Nostrand, New York.
10. Neyman, J. (1962). "Two Breakthroughs in the Theory of Statistical Decision Making." Review of the International Statistical Institute, 30 : 1, p. 11-27.
11. Robbins, H. (1951). "Asymptotic Subminimax Solutions of Compound Statistical Decision Problems." Proceedings of the Second Berkeley Symposium on Statistics and Probability, p. 131-148.
12. _____. (1955). "An Empirical Bayes Approach to Statistics." Proceedings of the Third Berkeley Symposium on Statistics and Probability, p. 157-163.

ANNEX 1

1. Annex 1.1: List of countries
Annex 1.1.1
2. Annex 1.2: List of countries
Annex 1.2.1
3. Annex 1.3: List of countries
Annex 1.3.1
4. Annex 1.4: List of countries
Annex 1.4.1
5. Annex 1.5: List of countries
Annex 1.5.1
6. Annex 1.6: List of countries
Annex 1.6.1
7. Annex 1.7: List of countries
Annex 1.7.1
8. Annex 1.8: List of countries
Annex 1.8.1
9. Annex 1.9: List of countries
Annex 1.9.1
10. Annex 1.10: List of countries
Annex 1.10.1

13. _____. (1963). "The Empirical Bayes Approach to Testing Statistical Hypotheses." Review of the International Statistical Institute, 31, p. 195-208.
14. _____. (1964). "The Empirical Bayes Approach to Statistical Decision Problems." Annals of Mathematical Statistics, 35, p. 1-20 .
15. Rosenblatt, M. (1956). "Remarks on Some Non-parametric Estimates of a Density Function." Annals of Mathematical Statistics, 27, p. 832-837.
16. Rutherford, J. R. and Krutchkoff, R. G. (1965). "The Parametric Empirical Bayes Approach to Statistical Hypothesis Testing and Point Estimation." Unpublished.
17. _____. (1965). "The Empirical Bayes Approach : Estimating the Prior Distribution." Submitted to Biometrika.
18. _____. (1965). "Empirical Bayes Estimation of the Prior and Posterior Distributions." To appear in Annals of Mathematical Statistics.
19. Samuel, E. (1963). "Asymptotic Solutions of the Sequential Compound Decision Problem." Annals of Mathematical Statistics, 34, p. 1079-1094.
20. _____. (1963). "An Empirical Bayes Approach to the Testing of Certain Parametric Hypotheses." Annals of Mathematical Statistics, 34, p. 1370-1385.
21. _____. (1964). "Convergence of the Losses of Certain Decision Rules for the Sequential Compound Decision Problem." Annals of Mathematical Statistics, 35, p. 1606-1621.
22. _____. (1965). "Sequential Compound Estimators." Annals of Mathematical Statistics, 36, p. 879-889.
23. Swain, D. D. (1965). "Bounds and Rates of Convergence for the Extended Compound Estimation Problem in the Sequence Case." Stanford University Technical Report No. 81.
24. Van Ryzin, J. R. (1966). "The Compound Decision Problem with $m \times n$ Finite Loss Matrix." Annals of Mathematical Statistics, 37, p. 412-424.
25. Von Mises, R. (1942), "On the Correct Use of Bayes Formula." Annals of Mathematical Statistics, 13, p. 156-165.

